AD-A111 938 OFFICE OF NAVAL RESEARCH ARLINGTON VA F/G 15/5
NAVAL RESEARCH LOGISTICS QUARTERLY. VOLUME 28, NUMBER 4,(U)
DEC 81

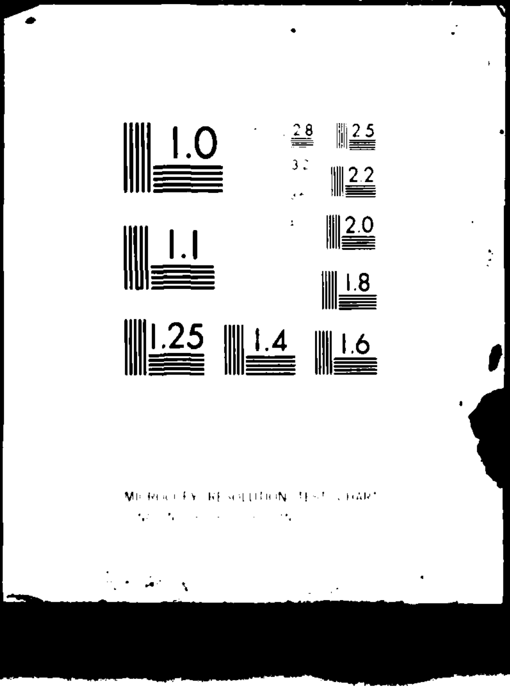
UNCLASSIFIED NL

1 OF 2
AD A
111938

1 OF 4

AD A

111938



1

NAVAL RESEARCH LOGISTICS QUARTERLY

DECEMBER 1981
VOL. 28, NO. 4

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $11.15 a year in the U.S. and Canada, $13.95 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

The views and opinions expressed in this Journal are those of the authors and not necessarily those of the Office of Naval Research.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, P-35 (Revised 1-74).

SOME PROPERTIES OF OPTIMAL CONTROL POLICIES FOR ENTRY TO AN M/M/1 QUEUE

Robert C. Rue

Department of Mathematical Sciences
USAF Academy, Colorado

Matthew Rosenshine

Department of Industrial and Management
Systems Engineering
The Pennsylvania State University
University Park, Pennsylvania

ABSTRACT

Customers served by an M/M/1 queueing system each receive a reward R but pay a holding cost of C per unit time (including service time) spent in the system. The decision of whether or not a customer joins the queue can be made on an individual basis or a social basis. The effect of increasing the arrival rate on the optimal policy parameters is examined. Some limiting results are also derived.

INTRODUCTION

The control of entry of customers to a queue is a virtual necessity in some cases and highly desirable in others. Rosenshine [6] presents a queue with state dependent service times for which no steady state distribution exists for any arrival rate and for which control of entry must eventually be exercised in the presence of continuing arrivals. Even queues for which steady state distributions exist can often be operated more effectively by controlling entry of customers, particularly those queues in which customers have a high expected waiting time.

Prior to the appearance of Naor's work [5] which dealt with optimal control of entry to a single-server queue, control was viewed primarily as a method of insuring adherence to some arbitrary externally imposed constraint, e.g., lim[illegible] queue size or expected waiting time, without regard for economic or efficient operation of the system. Yechiali [9] dealt with the same problem as Naor in a different, but equivalent, manner and provided what appears to be a more generally applicable procedure for determining the optimal control policy and the gain that can be achieved through its employment.

After the appearance of these early papers, work dealing with the application of optimal control of arrivals to more general queueing systems began to appear. Knudsen [2] and Yechiali [10] extended the optimal control solution to s servers, the Knudsen work considering a slightly more general cost structure than heretofore examined. A survey by Stidham and Prabhu [7] of the control of queueing systems work also pointed out that many of the papers surveyed had common considerations. Subsequently, Knudsen and Stidham [3] considered the case for which a new customer's net benefit for joining was a decreasing function of the number of customers already present in the system when the new customer arrived. Lippman and Stidham [4] examined the optimal control of a queue with state-dependent service rate, increasing with the number of customers in the system, and random rewards. They showed that regardless of the discounting policy and time horizon, the individual optimum policy admits customers to the system whenever the social optimum policy does (and perhaps other times as well). Stidham [8] extended this work to cover a GI/M/1 system. Recently, Johansen and Stidham [1] have shown that many of the properties of and relationships between the individual and social optimal control policies apply under very general conditions, e.g., dependent, nonidentically distributed, batch arrivals.

In this paper, we return to Naor's model and examine the sensitivity of its solution to changes in the arrival rate, λ. We do this for two cases, the individual optimum and the social optimum. In the former case, an arriving customer acts so as to maximize his own gain. In the latter, a customer's decision to join the queue or not is based on the effect of the decision on the gain rate (benefit or profit per unit time) of all customers wishing to use the queueing system. For the individual optimum case, the gain rate is shown to rise as λ increases and then fall because the control limit (maximum queue system size at which a customer will choose to join the queue) remains constant as λ changes. In the social optimum case, the maximum gain rate obtainable is proven to be a nondecreasing function of λ whereas the control limit which yields the maximum gain rate proves to be a nonincreasing function of λ. Thus, in all practical applications, as the arrival rate rises, it is possible to increase the gain rate at the expense of level of service (fraction of arrivals allowed to enter the queue).

A potential application of the model considered lies in the control of the arrival of aircraft to an airport. The control limit determined can be used in connection with an examination of actual traffic during busy periods to see when the traffic exceeds the control limit. The rescheduling of this excess traffic, delaying it prior to take-off, or the scheduling of this traffic into alternative terminals represent actions which will increase the overall gain rate for airline operations. At the same time, fuel expended while waiting to land will be reduced.

THE MODEL AND ITS SOLUTION

The model is an M/M/1 queueing model with arrivals occurring at a rate λ and having a service rate capability of μ. A cost structure is imposed on the operation of the system in which

a) each customer served receives a reward of R dollars, and

b) each unit of time a customer spends in the system costs him C dollars.

Each arriving customer is given the choice of joining the queue and receiving reward R and paying C per unit time in the system or of not joining and not paying or receiving any money. Customers are assumed to decide by comparing the expected net gain associated with each decision and choosing the action with the larger gain. (In case of a tie, the customer joins the queue.) This is the model considered by Naor [5] and Yechiali [9], although Yechiali allowed a general arrival distribution and a slightly more general cost structure.

Naor argues that all reasonable strategies lead to a finite capacity queue. He determines n_s, the capacity under self-optimization where each customer considers only his own expected net gain in deciding whether or not to join the queue. The expected net gain for joining is $R - (i + 1)\,C/\mu$, where i is the number of customers the arrival finds in the system. Joining the queue thus serves the self-interest of a customer if i is less than n_s, where n_s is defined by

$$R - (n_s + 1)C/\mu < 0 \leq R - n_s\,C/\mu.$$

This strategy leads to an M/M/1/n_s queueing system for which

$$n_s = [R\mu/C] \tag{1}$$

where [ ] indicates "the greatest integer in."

If each customer or an administrator acts to maximize the sum of the individual net benefits, the problem becomes the social optimum problem. Considering an infinite horizon without discounting, Naor sets up the following expected overall net benefit rate function:

$$g(n) = \lambda'(n)R - C\,L(n) \tag{2}$$

where $\lambda'(n)$ and $L(n)$ are, respectively, the effective arrival rate to and the expected number in the system when a maximum of n customers is allowed in the system. $g(n)$ is the expected net benefit rate or expected gain per unit time when a maximum of n customers is allowed in the system. The units of $g(n)$ are dollars per unit time.

Another form of (2) will be used in the remainder of this paper. This form is derived by noting first that

$$\lambda'(n) = \lambda(1 - \theta_n(n)) = \lambda \sum_{i=0}^{n-1} \theta_i(n)$$

where $\theta_i(n)$ is the stationary probability that there are i customers in the system when the balking point is n and second that

$$L(n) = \sum_{j=1}^{n} j\,\theta_j(n) = \frac{\lambda}{\mu} \sum_{j=1}^{n} j\,\theta_{j-1}(n) = \frac{\lambda}{\mu} \sum_{i=0}^{n-1} (i + 1)\,\theta_i(n).$$

Thus, (2) can be written as

$$g(n) = \sum_{i=0}^{n-1} \theta_i(n)\,\lambda\,(R - (i + 1)C/\mu). \tag{3}$$

Since the $\theta_i(n)$ represent the steady-state probabilities of a semi-Markov process, this equation represents the gain function for a semi-Markov decision process in which n must be decided. The optimum value of n can be determined by search or more conveniently by policy iteration. A derivation of this equation directly from the semi-Markov decision process formulation of this problem is found in [9].

SOME PROPERTIES OF OPTIMAL CONTROL POLICIES

The gain rate and control limit for the individual optimum and social optimum cases can be determined from Naor's results for a given value of λ. However, the arrival rate λ is subject to change. For example, λ grows as the usage of an airport grows. Thus, it is useful to know how the optimal control policy should be changed in response changes in system usage. It is these changes that we examine here.

An increase in λ, the arrival rate, has no effect on the solution to the individual optimum problem because a self-optimizing customer considers only his own expected net benefit in deciding whether or not to join the system. The balking point, n_s, is independent of λ as indicated by (1). Self-optimizing customers do not recognize any measure of the overall gain of all arrivals. They only consider their own expected gain which does not depend on λ. However, for any policy adopted by self-optimizing customers, the gain per unit time can be calculated. Sample results are shown in Table 1. Comment on these results is reserved until comparable social optimum results are presented.

TABLE 1 — *Gain Rates Implied by Individual Optimum for Various Values of* λ
$(R = 5,\ C = 2,\ \mu = 3)$

λ	n_s	g
0.1	7	.431
1.0	7	4.001
2.1	7	6.537
2.2	7	6.595
2.3	7	6.623
2.4	7	6.621
4.02	7	4.637
4.05	7	4.596
5.0	7	3.556
16.4	7	1.448
16.6	7	1.441
100.0	7	1.062

The solution of the social optimum problem is affected by an increase in λ. The effect of λ on the expected gain rate of the system is considered first.

THEOREM 1: g is a nondecreasing function of λ.

PROOF: Define $g(P_i, \lambda)$ to be the expected gain rate of the system under policy P_i when the arrival rate is λ. Let $g^*(\lambda)$ denote the maximum expected gain rate when the arrival rate is λ. Let $D_i(1)$ be the stationary probability that an arrival is allowed to join when the system contains i customers and, similarly, let $D_i(0) = 1 - D_i(1)$ be the stationary probability that an arrival must balk when i are in the system. Yechiali [9] shows that a deterministic control-limit policy is optimal so let $P_1 = \{D_i(1): D_i(1) = 1,\ i < n_0;\ D_i(1) = 0,\ i \geq n_0\}$ be such that $g^*(\lambda'') = g(P_1, \lambda'')$. Let $\lambda' > \lambda''$ and $P_2 = \{D_i(1): D_i(1) = \frac{\lambda''}{\lambda'},\ i < n_0;\ D_i(1) = 0,\ i \geq n_0\}$, where n_0 is the balking point from policy P_1. Under policy P_2, an arrival who finds fewer than n_0 customers in the system is allowed to join with probability λ''/λ' and is forced to balk with probability $1 - \lambda''/\lambda'$. Since no penalty is assessed for rejecting a customer, $g(P_2, \lambda') = g(P_1, \lambda'') = g^*(\lambda'')$. Finally, since $g^*(\lambda') \geq g(P_2, \lambda')$, $g^*(\lambda') \geq g^*(\lambda'')$.

Although the social optimum balking point can be determined for a given λ by policy iteration, the range of λ for which the balking point is a given integer can also be found as follows. Define $\{f(i)\}$ to be the sequence of expected rewards for joining, where $f(i) = R - C(i+1)/\mu$ is the expected reward if i customers are in the system. Then, from (3),

$$g(n) = \sum_{i=0}^{n-1} \theta_i(n)\,\lambda\, f(i). \tag{4}$$

Naor states that $g(n)$ is discretely unimodal in n. Thus, n_0 is such that $\Delta g(n_0+1) < 0 \leqslant \Delta g(n_0)$, where $\Delta g(n) = g(n) - g(n-1)$.

$$\begin{aligned}\Delta g(n) &= \sum_{i=0}^{n-1} \theta_i(n)\,\lambda\, f(i) - \sum_{i=0}^{n-2} \theta_i(n-1)\,\lambda\, f(i)\\ &= \sum_{i=0}^{n-2} (\theta_i(n) - \theta_i(n-1))\,\lambda\, f(i) + \theta_{n-1}(n)\,\lambda\, f(n-1).\end{aligned}$$

$\Delta g(n) \geq 0$ if

$$\theta_{n-1}(n)\,\lambda\, f(n-1) \geq \sum_{i=0}^{n-2} (\theta_i(n-1) - \theta_i(n))\,\lambda\, f(i). \tag{5}$$

If $\rho = \lambda/\mu \neq 1$, then (5) can be written as

$$(1-\rho)\rho^{n-1} f(n-1)/(1-\rho^{n+1}) \geq \sum_{i=0}^{n-2} \left\{\frac{(1-\rho)\rho^i}{(1-\rho^n)} - \frac{(1-\rho)\rho^i}{(1-\rho^{n+1})}\right\} f(i). \tag{6}$$

After formation of a common denominator on the right-hand side and division by $\{(1-\rho)/(1-\rho^{n+1})\}$, (6) becomes

$$f(n-1) \geq \frac{\{\rho(1-\rho)\}}{(1-\rho^n)} \sum_{i=0}^{n-2} \rho^i f(i).$$

Substitution of $\sum_{i=0}^{n-1} \rho^i$ for $(1-\rho^n)/(1-\rho)$ leads to $\Delta g(n) \geq 0$ if

$$f(n-1) \sum_{i=0}^{n-1} \rho^i \geq \rho \sum_{i=0}^{n-2} \rho^i f(i), \quad n \geqslant 2. \tag{7}$$

If $\rho = 1$, $\theta_i(n) = \dfrac{1}{n+1}$, so (5) becomes

$$f(n-1)/(n+1) \geq \sum_{i=0}^{n-2} \left\{\frac{1}{n} - \frac{1}{n+1}\right\} f(i)$$

or

$$n f(n-1) \geq \sum_{i=0}^{n-2} f(i). \tag{8}$$

Since the limit as $\rho \to 1$ of (7) is (8), (7) can be used for all values of ρ. Thus, for a given ρ, n_0 is the largest value of n such that (7) holds.

If μ is held constant, (7) can be used to find the range of λ for which n_0 is a given integer. (Note $\Delta g(n) < 0$ if $\geq$ is changed to $<$ in (7).) Finding the range of λ for which

$n_0 = 1$ requires finding λ such that $\Delta g(1) \geq 0$, but $\Delta g(2) < 0$. If $\Delta g(1) < 0$, $R < C/\mu$ and the system is trivial since customers never enter it. Thus, $\Delta g(1) \geq 0$ for all λ since $g(0) = 0$. Finding the range of λ over which $n_0 = j$, $j \geq 2$, requires finding λ such that $\Delta g(j) \geq 0$ and $\Delta g(j+1) < 0$. Note that solving (7) for λ such that $\Delta g(n) < 0$ or $\Delta g(n) \geq 0$ involves finding the roots of the $n-1$ degree polynomial resulting from $\Delta g(n) = 0$. Use of (7) is demonstrated by an example.

Suppose that the reward for service is $R = 5$, the cost per unit time in the system is $C = 2$ and the service rate of the single server is $\mu = 3$. $\{f(i)\} = \{R - C(i+1)/\mu\} = \{13/3, 11/3, 3, 7/3, 5/3, 1, 1/3, -1/3, \ldots\}$. (Note that $n_s = 7$.) First, find the range of λ for which $n_0 = 1$. $\Delta g(1) \geq 0$ for all λ. $\Delta g(2) < 0$ if from (7)

$$f(1)\sum_{i=0}^{1}\rho^i < \rho\sum_{i=0}^{0}\rho^i f(i) \quad \text{or} \quad 16.5 < \lambda.$$

Thus, $n_0 = 1$ if $\lambda > 16.5$. Now find the range of λ over which $n_0 = 2$. $\Delta g(2) \geq 0$ if $\lambda \leq 16.5$. $\Delta g(3) < 0$ if from (7)

$$f(2)\sum_{i=0}^{2}\rho^i < \rho\sum_{i=0}^{1}\rho^i f(i) \text{ or } -2\lambda^2 - 12\lambda + 81 < 0.$$

Solution of the above yields $\Delta g(3) < 0$ if $\lambda > 4.035$. Therefore, $n_0 = 2$ if $4.035 < \lambda \leq 16.5$. Further use of (7) yields $n_0 = 3$ if $2.1 < \lambda \leq 4.035$. Table 2 presents the solution found using policy iteration for several values of λ. These results confirm (7) and Theorem 1.

TABLE 2 — *Policy Iteration Results for Various Values of* λ
($R = 5$, $C = 2$, $\mu = 3$)

λ	n_0	g
0.1	7	0.431
1.0	5	4.003
2.1	4	6.944
2.2	3	7.128
4.02	3	8.993
4.05	2	9.011
16.4	2	10.998
16.6	1	11.010
100.00	1	12.621

Comparison of Tables 1 and 2 illustrates that $n_0 \leq n_s$ and that the gain rate associated with the individual optimum policy is no greater than the social optimum gain rate. Also, the gain rate associated with the individual optimum policy reaches a peak and then declines as λ increases. On the other hand, the social optimum gain rate is nondecreasing in λ as shown in Theorem 1.

The value of n_0, the forced balking point for the social optimum problem, appears to be nonincreasing as the arrival rate increases in Table 2. This result is established as Theorem 2.

THEOREM 2: n_0 is a nonincreasing function of λ.

PROOF: R and μ are assumed to be finite and C is assumed to be nonzero. Since Naor proved that g is discretely unimodal in n, n_0 is optimal if $\Delta g(n_0+1) < 0 \leq \Delta g(n_0)$. Let $n_0(\lambda)$ denote the optimal forced balking point when the arrival rate is λ. Let $n = n_0(\lambda'')$, where λ'' is a fixed value of λ and $n \geqq 2$. From (7)

$$\Delta g(n) \geqq 0 \text{ if}$$

$$f(n-1)\sum_{i=0}^{n-1}\rho^i \geqq \sum_{i=0}^{n-2}\rho^i f(i),$$

for $n \geqq 2$, which can be written as

$$\Delta g(n) \geqq 0 \text{ if}$$

$$f(n-1)(1+\rho+\ldots+\rho^{n-1}) \geqq \rho f(0)+\rho^2 f(1)+\ldots+\rho^{n-1}f(n-2). \tag{8}$$

As λ increases, ρ increases. Since $\{f(i)\}$ decreases as i increases, $f(0) > f(1) > \ldots > f(n-2) > f(n-1)$. Thus, as λ increases, the right-hand side of (8) increases faster than the left-hand side. This eventually leads to $\Delta g(n) < 0$ for λ greater than or equal to some $\lambda' > \lambda''$. Therefore, $n_0(\lambda') < n_0(\lambda'')$. Since $\Delta g(1) \geqq 0$ for all λ, continuation of the above argument leads to the existence of λ''' such that $n_0(\lambda) = 1$ for all $\lambda \geqq \lambda'''$. Thus, as λ increases, n_0 decreases until it becomes equal to one.

Since n_0 has been shown to be a nonincreasing function of λ and g has been shown to be a nondecreasing function of λ, the question of existence of the limits as $\lambda \to \infty$ of n_0 and g arises. From Theorem 2, n_0 decreases as λ increases until $n_0 = 1$. n_0 remains one for further increases in λ. Thus,

$$\lim_{\lambda\to\infty} n_0 = 1.$$

If $n_0 = 1$,

$$\phi_0(n_0) = \phi_0(1) = \frac{\mu}{\mu+\lambda} \text{ and } \phi_1(n_0) = \phi_1(1) = \frac{\lambda}{\mu+\lambda}.$$

From (4)

$$g(1) = \sum_{i=0}^{0}\phi_i(1)\lambda f(i) = \mu\lambda f(0)/(\mu+\lambda).$$

Dividing numerator and denominator of the right-hand side by λ and letting $\lambda \to \infty$, yields

$$\lim_{\lambda\to\infty} g(1) = \mu f(0). \tag{9}$$

For the example given previously $\lim_{\lambda\to\infty} g(1) = 13$.

Equation (9) can also be justified intuitively. If no queue is allowed to form, customers can only join if the system is empty. The expected reward for a joining customer is $f(0)$. The expected time between departures of customers from the system is $1/\mu$ plus the expected time between the completion of a service and the arrival of the next customer. In the limit, an arrival occurs at the instant a service is completed so the system is never empty. Thus, the expected gain rate of the system as $\lambda \to \infty$ is $\dfrac{f(0)}{1/\mu} = \mu f(0)$.

SUMMARY

The effect of an increasing arrival rate on the individually and socially optimal control policies for entry to an M/M/1 queue has been investigated. The results presented are useful in determining the sensitivity of the optimal control policies to changes in the customer arrival rate.

REFERENCES

[1] Johansen, S.G. and S. Stidham, Jr., "Control of Arrivals to a Stochastic Input-Output System," Industrial Engineering Report No. 78-1, North Carolina State University, Raleigh, NC (1978).

[2] Knudsen, N.C., "Individual and Social Optimization in a Multiserver Queue with a General Cost-Benefit Structure," Econometrica, *40*, 515-528 (1972).

[3] Knudsen, N.C. and S. Stidham, Jr., "Individual and Social Optimization in Birth-Death Congestion System with a General Cost-Benefit Structure," Industrial Engineering Report No. 76-8, North Carolina State University, Raleigh, NC (1976).

[4] Lippman, S.A. and S. Stidham, Jr., "Individual Versus Social Optimization in Exponential Congestion Systems," Operations Research, *25*, 233-247 (1977).

[5] Naor, P., "The Regulation of Queue Size by Levying Tolls," Econometrica, *37*, 15-24 (1969).

[6] Rosenshine, M., "Queues with State-Dependent Service Times," Transportation Research, *1*, 97-104 (1967).

[7] Stidham, S., Jr. and N.U. Prabhu, "Optimal Control of Queueing Systems," in *Mathematical Methods in Queueing Theory*, Lecture Notes in Economics and Mathematical Systems, *98*, 263-294 (Springer-Verlag, 1974).

[8] Stidham, S., Jr., "Socially and Individually Optimal Control of Arrivals to a GI/M/1 Queue," Management Science, *24*, 1598-1610 (1978).

[9] Yechiali, U., "On Optimal Balking Rules and Toll Charges in the GI/M/1 Queueing Process," Operations Research, *19*, 349-370 (1971).

[10] Yechiali, U., "Customer's Optimal Joining Rules for the GI/M/s Queue," Management Science, *18*, 434-443 (1972).

THE CONSTRUCTION OF AN OPTIMAL DISTRIBUTION OF SEARCH EFFORT

Ingo Wegener

Faculty of Mathematics
Bielefeld University
Federal Republic of Germany

ABSTRACT

Suppose one object is hidden in the k-th of n boxes with probability $p(k)$. We know the probability $q(t, k)$ of detecting the object if it is hidden in box k and we expend effort t searching box k. Our aim is to minimize the expected search effort of a successful search. Previously this problem has been solved only under the assumption that the functions $q(\cdot, k)$ are concave. We prove, without concavity assumptions, the existence of an optimal distribution of search effort and give a procedure for its construction.

INTRODUCTION

An absent-minded person repeatedly loses his glasses in his own house. In order to search optimally for the glasses he estimates the following data: $p(k)$ for $k \in B = \{1, \dots, n\}$, the *a*-priori probability that he has lost his glasses in the k-th room, and $q(t, k)$ for $t \in \mathbf{R}_0^+$ and $k \in B$, the conditional probability that he finds the glasses if he has lost them in room k and he searches this room for a time period of length t. We assume that $q(\cdot, k)$ is a probability distribution function, i.e., it is increasing and right-continuous.

More important research areas where we can apply this problem are the following: the search for a defect in a large system, the search for oil or a mineral in different regions and the search for a sunken ship where we partition the ocean into cells. In the following we discuss the search for an object hidden in one of a finite number, n, of boxes.

Now we have to decide how we should search for the hidden object. A search procedure (t_i, k_i) for $i \in \mathbb{N}$, $t_i \in \mathbf{R}^+$ and $k_i \in B$ prescribes the following behaviour: search box k_1 for t_1 hours, then search box k_2 for t_2 hours and so on until the object has been found.

These search strategies are appropriate if only one person searches for the object. But while searching for a sunken ship there may be many boats or planes. In this case, we can specify that one half of them searches cell 1, another third searches cell 5 and so on. We assume that the number of boats and planes is so large that it is a reasonable approximation to assume that we can divide the search effort in an arbitrary fashion.

A strategy is a function $s : \mathbf{R}_0^+ \times B \rightarrow \mathbf{R}_0^+$ where $s(t, k)$ specifies how much of the search effort t has to be spent in box k. Therefore, $\sum_{k \in B} s(t, k) = t$ and $s(\cdot, k)$ is increasing. By $Q(s, t)$ we denote the probability that we find the object when we spend search effort t. Obviously,

$$Q(s, t) = \sum_{k \in B} p(k) q(s(t, k), k)$$

and

$$Q(s, \infty) = 1 - \lim_{t \rightarrow \infty} Q(s, t)$$

is the probability that we never find the object if we use s. The function $Q(s, \cdot)$ is the probability distribution function of the random variable $Y(s)$ which measures the search effort which is necessary to find the object by strategy s. The mean of $Y(s)$ is the expected search effort of s and will be denoted by $E(s)$. It is a well known result of probability theory that

$$E(s) = \int_{[0, \infty]} t \, dQ(s, \cdot) = \int_{[0, \infty)} (1 - Q(s,t)) dt.$$

A strategy s is called optimal iff its expected search effort is minimal.

Gilbert [3] and Kisi [5] determined optimal strategies for some definite concave functions, $q(\cdot, k)$. Onaga [6] gave a solution of the problem under the assumption that all $q(\cdot, k)$ are concave. As we shall see in this paper the assumption of Onaga is a considerable simplification of the problem. Here we have to mention that these three authors tried to solve a more general and much more difficult search problem. They assumed that one has to pay some extra costs, called switch costs, if one changes the place of search. Stone [7] proved the results of Onaga using new methods which we present in Section 2.

In Section 1 we give a necessary and sufficient condition for the existence of a strategy with finite expected search effort. In Sections 2 and 3 we repeat briefly the solution of the problem under the assumption that the functions $q(\cdot, k)$ are concave. Later, we shall use these results for the solution of the general problem.

In Section 2 we assume that our funds are limited by $T \in \mathbf{R}_0^+$. The best we can do is to distribute this search effort in such a way that the probability of finding the object is maximal. A T-admissible allocation is a function $a : B \rightarrow \mathbf{R}_0^+$ where $\sum_{k \in B} a(k) = T$. The allocation a prescribes that we spend $a(k)$ for the search of box k. The probability of success of a becomes

$$P(a) = \sum_{k \in B} p(k) q(a(k), k).$$

We construct a T-optimal allocation. In Section 3 we use this result and construct an optimal strategy.

In the following sections we do not assume that $q(\cdot, k)$ is concave. Our main result is the following. Let $\bar{q}(\cdot, k)$ be the smallest concave function nowhere smaller than $q(\cdot, k)$. The optimal strategy of Section 3 for the search problem where $q(\cdot, k)$ is replaced by $\bar{q}(\cdot, k)$, is also optimal for the problem with the functions $q(\cdot, k)$. In Section 4 we show how we may improve strategies and in Section 5 we prove the main result.

Finally (Section 6) we compare our results with two similar search problems. First is the well-known problem where one is not allowed to divide the search effort in an arbitrary fashion, but where the cost of the j-th search of box k and the probability of overlooking the object during this search (even if one searches the right box) is given. Second we consider the search problem where the hidden objec. is a point of $\mathbf{R}^n$.

1. THE EXISTENCE OF A STRATEGY WITH FINITE EXPECTED SEARCH EFFORT

Without loss of generality, we assume that $p(k) > 0$ for all k. Let $E(s|k)$ be the expected search effort of strategy s if the object is hidden in box k. Since

$$E(s) = \sum_{k \in B} p(k) E(s|k),$$

$E(s)$ is finite iff $E(s|k)$ is finite for all k. If the object is hidden in box k the best we can do is to search only box k. That means s^k defined by $s^k(t, k) = t$ and $s^k(t, k') = 0$ for $k' \neq k$ is optimal in that case. Therefore,

$$E(s|k) \geqslant E(s^k|k) = \int_{[0,\infty)} (1 - q(t, k))dt$$

and the strategy s may have finite expected search effort only if $E(s^k|k)$ is finite for all k.

THEOREM 1: There exists a strategy with finite expected search effort iff $\int_{[0,\infty)} (1 - q(t, k))dt < \infty$ for all k.

PROOF: We have already proved the only-if part. Now we may assume that

$$E(s^k|k) = \int_{[0,\infty)} (1 - q(t, k))dt < \infty$$

for all k. Let s defined by $s(t,k) = t/n$ be the strategy which distributes the search effort equally to all boxes. Obviously,

$$E(s|k) = nE(s^k|k) < \infty$$

and

$$E(s) = \sum_{k \in B} p(k) E(s|k) < \infty. \qquad \text{Q.E.D.}$$

In the following we assume the existence of a strategy with finite expected search effort.

2. OPTIMAL ALLOCATIONS FOR THE CONCAVE CASE

In this chapter we assume that the functions $q(\cdot, k)$ are concave. Charnes and Cooper [2] solved this problem for some definite concave functions $q(\cdot, k)$. A general solution has been obtained by Stone [7]. We cite his results and then we choose a special t-optimal allocation which will be useful for our considerations for nonconcave functions, $q(\cdot, k)$.

Stone used optimization techniques involving Lagrange multipliers. These basic techniques for the theory of search have been presented by Wagner and Stone [8]. We need the following function l defined by

$$l(k,\lambda,t) := p(k)q(t,k) - \lambda t \text{ for } k \in B,\ \lambda \in \mathbf{R}_0^+ \text{ and } t \in \mathbf{R}_0^+$$

which is called the pointwise Lagrangian. Stone [7] has shown that a t-admissible allocation a_t is t-optimal if for some $\lambda \geqslant 0$

$$l(k,\lambda,a_t(k)) = \max\{l(k,\lambda,z) \mid z \in \mathbf{R}_0^+\} \text{ for all } k \in B.$$

The problem of maximizing the pointwise Lagrangian is a problem of calculus. The following lemma states some easy facts.

LEMMA 1: Let $q(\cdot,k)$ be a concave probability distribution function on $\mathbf{R}_0^+$. $q'(t,k) := \lim_{h\to 0, h>0} h^{-1}(q(t+h,k) - q(t,k))$ is well defined and nonnegative for $t \geqslant 0$. It may happen that $q'(0,k) = \infty$. $q'(\cdot,k)$ is decreasing and right-continuous. $'q(t,k) := \lim_{h\to 0, h<0} h^{-1}(q(t+h,k) - q(t,k))$ is well defined and nonnegative for $t > 0$. $'q(\cdot,k)$ is decreasing and left-continuous.

$$q'(t,k) \leqslant {}'q(t,k) \text{ for all } t \text{ and } k.$$

By these properties it is easy to see that a_t maximizes the pointwise Lagrangian for some $\lambda \geqslant 0$ if

$$p(k)q'(a_t(k),k) \leqslant \lambda \text{ and } p(k)'q(a_t(k),k) \geqslant \lambda \text{ if } a_t(k) > 0$$

and

$$p(k)q'(a_t(k),k) \leqslant \lambda \text{ if } a_t(k) = 0.$$

We now choose for each $t \geqslant 0$ a definite t-optimal allocation in the following way. Let

$$T^{>}(L,k) := \sup\{t \in \mathbf{R}_0^+ \mid p(k)q'(t,k) > L\}$$

;and

$$T^{\geqslant}(L,k) := \sup\{t \in \mathbf{R}_0^+ \mid p(k)q'(t,k) \geqslant L\}$$

for $L \in \bar{\mathbf{R}}_0^+$ and $k \in B$.

Obviously, $T^{>}(\infty,k) = T^{\geqslant}(\infty,k) = 0$ and $T^{\geqslant}(0,k) = \infty$. For each $t \in \mathbf{R}_0^+$ we may define an allocation a_t^* with the following properties: a_t^* is t-admissible; there is an $L(t) \in \bar{\mathbf{R}}_0^+$ and a box $j(t)$ so that $a_t^*(k) = T^{\geqslant}(L(t),k)$ for $k < j(t)$, $a_t^*(k) = T^{>}(L(t),k)$ for $k > j(t)$ and $a_t^*(j(t)) \in [T^{>}(L(t),k), T^{\geqslant}(L(t),k)]$.

Again, by lemma 1 and our results above, a_t^* maximizes the pointwise Lagrangian for $\lambda = L(t)$ and is therefore t-optimal.

THEOREM 2: a_t^* is an optimal allocation of the search effort t.

3. OPTIMAL STRATEGIES FOR THE CONCAVE CASE

By the definition of a_t^* we may conclude that $a_t^*(k)$ is increasing as a function of t. Thus, s^*, defined by $s^*(t,k) := a_t^*(k)$, is a strategy. Again, by the results of Stone [7], s^* is optimal. This may be seen directly, also. Let s be a strategy. Then a_t defined by $a_t(k) := s(t,k)$ is a t-admissible allocation. By theorem 2

$$\begin{aligned} E(s^*) &= \int_{[0,\infty)} (1 - Q(s^*, t))dt = \int_{[0,\infty)} (1 - P(a_t^*))dt \\ &\leqslant \int_{[0,\infty)} (1 - P(a_t))dt = \int_{[0,\infty)} (1 - Q(s, t))dt = E(s). \end{aligned}$$

THEOREM 3: s^* is an optimal strategy.

4. AN IMPROVEMENT RULE

For each search problem which is defined by n, p and q we define the dual problem by $\bar{n}$, $\bar{p}$ and $\bar{q}$ where $\bar{n} := n$, $\bar{p} := p$ and $\bar{q}(\cdot, k)$ is the smallest concave function nowhere smaller than $q(\cdot, k)$. By elementary analysis one can prove that for each $\bar{q}(\cdot, k)$ there is a (perhaps empty) set of disjoint intervals (t_1, t_2) so that outside of these intervals $\bar{q}(\cdot, k)$ and $q(\cdot, k)$ coincide, while $\bar{q}(\cdot, k)$ is on each interval (t_1, t_2) a linearly and strictly increasing function larger than $q(\cdot, k)$. We denote these intervals by (t_1, t_2, k) to mark the appertaining box. The set of all these triples will be called I.

Let us investigate the optimal strategy s^* for the dual problem which we have constructed in Sections 2 and 3. The search effort for box k increases from t_1 to t_2 within a time interval of length $t_2 - t_1$ if $(t_1, t_2, k) \in I$. That means s^* prescribes that we have to search for a time period of length $t_2 - t_1$ only for box k and the search effort for k increases during this time period from t_1 to t_2. This will be called the (t_1, t_2, k) - property.

In the following we shall see that it would not be sufficient to use an arbitrary optimal strategy for the dual problem instead of an optimal strategy fulfilling the (t_1, t_2, k) - property for each $(t_1, t_2, k) \in I$. We explain this fact in the following way. If we increase the search effort for the box k from t_1 to t_2, then our profit is increased by $p(k)q(t_2, k) - p(k)q(t_1, k)$ while our search effort increases by $t_2 - t_1$. The quotient of profit and cost may be called efficiency. If $(t_1, t_2, k) \in I$ then it is less efficient to increase the search effort for box k from t_1 to $t_1 + h$ (for $h < t_2 - t_1$) than to increase it from $t_1 + h$ to t_2. Then, using the ideas for the solution of discrete search problems (Wegener [9]), we may believe that good strategies have the (t_1, t_2, k) - property. The following lemma makes these ideas precise.

LEMMA 2: For each strategy s and for each $(t_1, t_2, k) \in I$ there exists a strategy s' with the (t_1, t_2, k) - property which is at least as good as s.

PROOF: If s has infinite expected search effort the assertion is obviously correct. Otherwise, we may choose $T_1 \in \mathbf{R}_0^+$ and $T_2 \in \bar{\mathbf{R}}_0^+$ so that $s(T_1, k) = t_1$ and $s(T_2, k) = t_2$. The case where we have to choose $T_2 = \infty$ is a little more complicated. We give our proof for the case $T_2 < \infty$ and we add some comments as to how we change the proof if $T_2 = \infty$.

Let H be the set of all functions $h: [0, T_2 - T_1] \to [0, t_2 - t_1]$ where $0 \leqslant h(x) - h(y) \leqslant x - y$ if $x > y$, $h(0) = 0$ and $h(T_2 - T_1) = t_2 - t_1$.

In the following we restrict ourselves to strategies $\bar{s}$ where $\bar{s}(t, \cdot) = s(t, \cdot)$ for $t \leqslant T_1$ or $t \geqslant T_2$. For such a strategy $\bar{s}$ and $x \in [0, T_2 - T_1]$, let $h(x) := \bar{s}(T_1 + x, k) - \bar{s}(T_1, k)$. Obviously, $h \in H$.

On the other hand, we may define for $h \in H$ a corresponding strategy in the following way:

$$\bar{s}(t, \cdot) := s(t, \cdot) \text{ if } t \leqslant T_1 \text{ or } t \geqslant T_2,$$

$$\bar{s}(t, k) := s(T_1, k) + h(t - T_1) \text{ if } T_1 \leqslant t \leqslant T_2 \text{ and}$$

$$\bar{s}(t, k') := s(T(t), k') \text{ if } T_1 \leqslant t \leqslant T_2 \text{ and } k' \neq k$$

where $T(t)$ is chosen so that $\sum_{k'' \in B} \bar{s}(t, k'') = t$. This definition means that we treat all boxes, $k', k'' \neq k$, in the same way. If for s there is a point of time T where $s(T, k') = t'$ and $s(T, k'') = t''$ then there is a point of time T' where $\bar{s}(T', k') = t'$ and $\bar{s}(T', k'') = t''$.

Let H^* be the following subset of H. For $h \in H^*$ there exists $y \in [0, T_2 - T_1 - (t_2 - t_1)]$ so that $h(y) = 0$ and $h(y + t_2 - t_1) = t_2 - t_1$. Then the corresponding strategy s does not spend any effort for box k in the intervals $[T_1, y]$ and $[y + t_2 - t_1, T_2]$ while we search only box k from the point of time y to $y + t_2 - t_1$. This function, h, will now be denoted by h_y. We conclude that a strategy corresponding to a function of H has the (t_1, t_2, k) - property iff the corresponding function h is an element of H^*.

But at first we consider another subset H_m of H and the corresponding set of strategies S_m. $h \in H_m$ iff we may divide the interval $[0, T_2 - T_1]$ to 2m parts so that h is constant on m parts of length $(T_2 - T_1 - (t_2 - t_1))/m$ each while on the other m parts of length $(t_2 - t_1)/m$ each h is linearly increasing with slope 1.

If $T_2 = \infty$ the first m parts are replaced by m^2 parts of length $(t_2 - t_1)/m$ each, and one part of infinite length, which of course is always the last of all $m^2 + m + 1$ parts. Since obviously H_m is a finite set there exists a best of all strategies in S_m.

In order to compute a best strategy in S_m we have to decide how to arrange the m periods when we search box k and the m periods (if $T_2 = \infty$ m^2 periods) when we search other boxes. We assume that we are not able to find the hidden object during a period of search, but that we find the object at the end of a period with the same probability which formerly was the probability of finding the object during the whole period. By this assumption the expected search effort of all strategies, $\bar{s} \in S_m$, is increased by the same amount. Therefore, a best of all strategies in S_m for the new situation is also the best strategy in S_m for the original situation.

Now we are in a situation where we have to arrange optimally m searches of box k and m other searches (if $T_2 = \infty$, m^2 other searches) and where we may find the object only at the end of a search. This is the situation of a discrete search problem which has been solved by Wegener [9]. By the results of that paper the searches of box k belong together since $(t_1, t_2, k) \in I$. Therefore, s_m^*, the best strategy in S_m, prescribes the m searches of box k one after another. Hence, the corresponding function is of the form h_{y_m} for an appropriate y_m.

Now we like to approximate the given strategy s by a strategy $s_m \in S_m$. By elementary analysis it is easy to define s_m in such a way that the event that the search effort for box $k' \in B$ reaches $t \in \mathbf{R}_0^+$ if we use s_m has a delay of at most

$$\max\{(t_2 - t_1)/m, (T_2 - T_1 - (t_2 - t_1))/m\} \leqslant (T_2 - T_1)/m$$

compared with the same event for the given strategy s. Therefore,

$$E(s_m^*) \leqslant E(s_m) \leqslant E(s) + (T_2 - T_1)/m \text{ and}$$

$$\overline{\lim_{m\to\infty}}\ E(s_m^*) \leqslant E(s).$$

If $T_2 = \infty$ the delay is at most $(t_2 - t_1)/m$ plus that search effort for box k which is prescribed by s after the point of time where we now begin with the period of infinite length. This point of time is $T_1 + (t_2 - t_1) + m(t_2 - t_1) \to \infty$ for $m \to \infty$. Since $\lim\limits_{t\to\infty} s(t, k) = t_2 < \infty$ the second term of the delay tends to 0, too. We obtain, in the same way as above, the result $\overline{\lim\limits_{m\to\infty}}\ E(s_m^*) \leqslant E(s)$.

For s_m^* we have defined the corresponding function h_{y_m}. Since $y_m \in [0, T_2 - T_1 - (t_2 - t_1)]$, there is a subsequence $y_{m'}$ of y_m which converges to $y' \in [0, T_2 - T_1 - (t_2 - t_1)]$. If $T_2 = \infty$ we may easily prove that the corresponding strategy s_0 of h_0 is better than the corresponding strategy of h_y if $y \geqslant A$ for an appropriate $A < \infty$. Since $s_0 \in S_m$ for all m we conclude that $y_m \in [0, A]$ and, for that reason, y_m has a convergent subsequence.

We define s' as the strategy corresponding to $h_{y'}$. s' has the (t_1, t_2, k) - property and

$$E(s') \leqslant \overline{\lim_{m\to\infty}}\ E(s_{m'}^*) \leqslant \overline{\lim_{m\to\infty}}\ E(s_m^*) \leqslant E(s). \qquad \text{Q.E.D.}$$

REMARK: There is at most one triple (t_1', t_2', k') so that s has the (t_1', t_2', k') - property and s' has not. But this interval has been broken into only 2 parts. Again, by an easy application of the results [9] we obtain a strategy s'' at least as good as s which has the (t_1, t_2, k) - property and the (t_1^*, t_2^*, k^*) - property for all triples for which s has this property.

We note that we have not yet proved that we may restrict ourselves to strategies which have the (t_1, t_2, k) - property for all triples of I. If we apply lemma 2 and the remark for all triples of I, one after another, we have to apply it (in certain cases) for infinitely many times. It is not obvious that the resulting sequence of strategies converges to a strategy which is at least as good as the given strategy.

5. OPTIMAL STRATEGIES FOR THE GENERAL CASE

By S^* we denote the set of all strategies which have the (t_1, t_2, k) - property for all triples of I.

LEMMA 3: The optimal strategy s^* which we have constructed for the dual problem is the best strategy of S^* for the original problem.

PROOF: Let $s \in S^*$. For each $(t_1, t_2, k) \in I$ there exist $T_1, T_2 \in \mathbf{R}_0^+$ so that $T_2 - T_1 = t_2 - t_1$, s searches only box k during the period of time $[T_1, T_2]$ and the search effort for k increases during this period of time from t_1 to t_2. For $t \in [T_1, T_2]$

$$\bar{Q}(s, t) - Q(s, t) = p(k)(\bar{q}(t_1 + t - T_1, k) - q(t_1 + t - T_1, k))$$

is independent of $s \in S^*$ and for all other $t \in \mathbf{R}_0^+$

$$\bar{Q}(s, t) - Q(s, t) = 0.$$

Thus, $E(s) - \bar{E}(s)$ is independent of $s \in S^*$ if $\bar{E}(s) < \infty$. ($\bar{E}$ and $\bar{Q}$ indicate that we use s for the dual problem.) If $\bar{E}(s) = \infty$ we conclude $E(s) = \infty$ since $E(s) \geqslant \bar{E}(s)$. By definition $s^* \in S^*$ and by Theorem 3 s^* is optimal for the dual problem. Combining our results we have proved the assertion. Q.E.D.

LEMMA 4: The optimal strategy s^* which we have constructed for the dual problem has finite expected search effort for the original problem.

PROOF: We have assumed that there exists a strategy s with finite expected search effort. Thus,

$$\int_{[0,\infty)} (1 - \bar{Q}(s^*, t))dt = \bar{E}(s^*) \leqslant \bar{E}(s) \leqslant E(s) < \infty.$$

By definition

$$E(s^*) = \int_{[0,\infty)} (1 - Q(s^*, t))dt = \int_{[0,\infty)} (1 - \bar{Q}(s^*, t))dt$$
$$+ \int_{[0,\infty)} (\bar{Q}(s^*, t) - Q(s, t))dt$$

where the first term is finite.

Since $s^* \in S^*$ we may divide $[0,\infty)$ into intervals where $\bar{Q}(s^*, t) = Q(s^*, t)$ and intervals (corresponding to the triples of I) where $\bar{Q}(s^*, t)$ is linearly increasing and $Q(s^*, t) < \bar{Q}(s^*, t)$. We may reflect the area between $Q(s^*, t)$ and $\bar{Q}(s^*, t)$ at the straight line $\bar{Q}(s^*, t)$. Since the reflected areas are disjoint and lie between $\bar{Q}(s^*, t)$ and the constant function 1, the second term above is not larger than the first term. Thus $E(s^*) < \infty$. Q.E.D.

THEOREM 4: The optimal strategy s^* which we have constructed for the dual problem is optimal for the original problem, too.

PROOF: If the statement of the theorem is false there exists a strategy s' with finite expected search effort so that $\epsilon := E(s^*) - E(s') > 0$. Without loss of generality we may assume that $\lim_{t\to\infty} s^*(t, k) = \lim_{t\to\infty} s'(t, k)$ for all k. We prove the theorem by defining a strategy $s \in S^*$ which is better than s^*. This is a contradiction to Lemma 3. Starting with s' we shall define s step by step.

STEP 1: We choose $T \in \mathbf{R}_0^+$ so that

$$\int_{(T,\infty]} t dQ(s^*, \cdot) < \epsilon/2$$

(this is possible since $E(s^*) < \infty$ by lemma 4) and so that $s^*(T,k) \leqslant t_1$ or $s^*(T,k) \geqslant t_2$ for each triple $(t_1, t_2, k) \in I$ (this is possible since $s^* \in S^*$). Let I' be the set of triples where $s^*(T, k) \geqslant t_2$.

Let $s''(t, k) := s^*(t, k)$ for $t \geqslant T$ and for $t \leqslant T$ we imitate s'. For $t \in [0, T]$ we define

$$s''(t, k) := \min\{s^*(T, k), s'(t', k)\}$$

where we choose t' in such a way that $\sum_{k \in B} s''(t, k) = t$.

$$E(s'') = \int_{[0,T]} t dQ(s'', \cdot) + \int_{(T,\infty]} t dQ(s'', \cdot).$$

The event that the search effort for box k reaches $t \leqslant s^*(T, k)$ if we use s'', happens not later than the same event if we use s'. The first term of the above equation for $E(s'')$ is equal to the following value. If we find the object in box k with search effort $t \leqslant s^*(T, k)$, we have to pay the whole search effort which we have spent using s'' which is not more than the corresponding search effort if we use s'. If we find the object after we have spent more than $s^*(T, k)$ for box k, we have to pay nothing which is also not more than the corresponding search effort for s'. By these considerations the first term is not larger than $E(s')$. The second term equals $\int_{(T,\infty]} t dQ(s^*, \cdot)$ and is smaller than $\epsilon/2$. Thus,

$$E(s'') < E(s') + \epsilon/2.$$

STEP 2: Since $\sum_{(t_1,t_2,k) \in I'} (t_2 - t_1) \leqslant T$ we may define $I'' \subseteq I'$ so that I'' contains only finitely many elements and so that

$$\sum_{(t_1,t_2,k) \in I' - I''} (t_2 - t_1) < \epsilon/2.$$

We apply Lemma 2 and the Remark to s'' for each triple of I'' one after another. The resulting strategy is called $\bar{s}$ and

$$E(\bar{s}) \leqslant E(s'') < E(s') + \epsilon/2.$$

STEP 3: $\bar{s}$ has the (t_1,t_2,k) - property for all triples of I except perhaps the triples of $I^* := I' - I''$. But the length of all intervals which belong to triples of I^* is altogether very small.

We define a sequence of strategies $s_0, \ldots, s_n$. Let $s_0 := \bar{s}$. If s_{k-1} is already defined we define s_k in the following way. For each triple $(t_1, t_2, k) \in I^*$ we choose the least point of time t' where $s_{k-1}(t', k) = t_1$ and a point of time t'' where $s_{k-1}(t'', k) = t_2$. Outside the intervals (t', t'') we do not change s_{k-1}. Inside these intervals we change s_{k-1} to the strategy which corresponds to h_0 (see the proof of Lemma 2). That means we search at first only box k for a period of time of length $t_2 - t_1$, and for $t \in [t' + t_2 - t_1, t'']$ let $s_k(t, k) = t_2$ and $s_k(t, k') = s_{k-1}(T(t), k')$ for $k' \neq k$ and an appropriate $T(t)$ so that $\sum_{k'' \in B} s_k(t, k'') = t$. By the choice of t' we don't destroy any (t_1, t_2, k) - property. Let $s := s_n$.

By our procedure s is obviously an element of S^*. The event that the search effort for box k reaches t if we use s happens at most by

$$\sum_{(t_1,t_2,k) \in I^*} (t_2 - t_1) < \epsilon/2$$

later than the same event for $\bar{s} = s_0$. Thus,

$$E(s) < E(\bar{s}) + \epsilon/2 < E(s') + \epsilon \leqslant E(s^*)$$

and $s \in S^*$. Because of this contradiction to Lemma 3 we have proved the Theorem. Q.E.D.

6. CONCLUSION

In this paper we have solved a search problem where the search space is discrete but where time is continuous, i.e., at each point of time we may find the object. Here we like to compare our results with results for some similar search problems.

Many authors have investigated a discrete search problem where one may find the object only at the end of some definite searches. This search problem has been solved by Wegener [9]. These results were useful for the solution of our search problem since we used these results especially in Sections 4 and 5. In this discrete search problem there does not exist always an optimal strategy while in this paper we could prove the existence of an optimal strategy in each situation. Finally, we like to mention that Kadane and Simon [4] obtained for discrete search problems similar results in the situation where only finitely many searches have a positive probability of success.

Arkin [1] solved a search problem which may be called the continuous version of our search problem. The hidden object is a point of $\mathbf{R}^n$. The a-priori probability is given by a density function $p:\mathbf{R}^n \rightarrow \mathbf{R}_0^+$. $q(\cdot, x)$ is for each point $x \in \mathbf{R}^n$ the probability distribution function for the detection of an object at point x. (We omit the measure theoretical assumptions.) A strategy s has to prescribe how we distribute the search effort: $s:\mathbf{R}_0^+ \times \mathbf{R}^n \rightarrow \mathbf{R}_0^+$ so that $s(\cdot, x)$ is increasing and

$$\int_{\mathbf{R}^n} s(t, x)dx = t$$

for all t. The probability of detecting the object until the point of time t amounts to

$$Q(s, t) = \int_{\mathbf{R}^n} p(x)q(s(t, x), x)dx.$$

The aim is to minimize the expected search effort of a successful search. At first Arkin solved the dual search problem $(\overline{p}, \overline{q})$ where $\overline{p} := p$ and $\overline{q}(\cdot, x)$ is the smallest concave function nowhere smaller than $q(\cdot, x)$. The methods of Arkin and our methods of Sections 2 and 3 resemble each other.

But if we drop the assumption that the functions $q(\cdot, k)$ and $q(\cdot, x)$ are concave the solution of the continuous search problem is much easier than the solution of our problem. Since each point of $\mathbf{R}^n$ has the measure 0, $s(\cdot, x)$ is not necessarily continuous while for our search problem $s(\cdot, k)$ is always continuous. This is the reason why there exists an optimal strategy s (for the continuous dual problem) so that $s(t, x) \notin (t_1, t_2)$ for each triple $(t_1, t_2, x) \in I$. That means that the search effort jumps over the crucial intervals. This optimal strategy has the same expected search effort whether we use it for the dual or the original problem. Since $\overline{q} \geqslant q$ the expected search effort of each strategy is for the original problem not smaller than for the dual problem. Thus, this optimal strategy (for the dual problem) is optimal for the original problem, too. This result is similar to our main result. Stone ([7], Theorem 2.4.6) gives a more elegant proof for the results of Arkin. Finally, we like to state an important difference between the continuous problem and our problem. We call a strategy s uniformly optimal if for each $t \geqslant 0$ the allocation a_t where

$$a_t(k) := s(t, k) \text{ resp. } a_t(x) := s(t, x)$$

is t-optimal. While for the continuous problem the optimal strategy is even uniformly optimal there does not always exist a uniformly optimal strategy for our problem if the functions $q(\cdot, k)$ are not concave. For an example see Stone [7] (Example 2.2.9).

Again Stone [7] gives an example (2.4.8) that even for the continuous search problem there may not exist a uniformly optimal strategy if the functions $q(\cdot, x)$ do not have to be right-continuous.

Altogether, our results together with the results of Arkin [1] and Stone [7], guarantee the existence of strategies which minimize the expected search effort for any "reasonable" search problem involving continuous effort.

REFERENCES

[1] Arkin, V.I., "A Problem of Optimum Distribution of Search Effort," Theory of Probability and Its Applications, *9*, 674-680 (1964).

[2] Charnes, A. and W.W. Cooper, "The Theory of Search-Optimum Distribution of Search Effort," Management Science, *5*, 44-50 (1958).

[3] Gilbert, E.N., "Optimal Search Strategies," SIAM, *7*, 413-424 (1959).

[4] Kadane, J.B. and H.A. Simon, "Optimal Strategies for a Class of Constrained Sequential Problems," Annals of Statistics, *5*, 237-255 (1977).

[5] Kisi, T., "On an Optimal Searching Schedule," Journal of the Operations Research Society of Japan, *8*, 53-65 (1965).

[6] Onaga, K., "Optimal Search for Detecting a Hidden Object," SIAM, *20*, 298-318 (1971).

[7] Stone, L.D., *Theory of Optimal Search* (Academic Press, New York, 1975).

[8] Wagner, D.H. and L.D. Stone, "Necessity and Existence Results on Constrained Optimization of Separable Functionals by a Multiplier Rule," SIAM Journal on Control, *12*, 356-372 (1974).

[9] Wegener, I., "The Discrete Sequential Search Problem with Nonrandom Cost and Overlook Probabilities," Mathematics of Operations Research, *5*, 373-380 (1980).

A VERSATILE SCHEME FOR RANKING THE EXTREME POINTS OF AN ASSIGNMENT POLYTOPE*

Mokhtar S. Bazaraa

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia

Hanif D. Sherali

School of Industrial Engineering and Operations Research
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

ABSTRACT

A cutting plane scheme embedded in an implicit enumeration framework is proposed for ranking the extreme points of linear assignment problems. This method is capable of ranking any desired number of extreme points at each possible objective function value. The technique overcomes storage difficulties by being able to perform the ranking at any particular objective function value independently of other objective values. Computational experience on some test problems is provided.

1. INTRODUCTION

The ranking of extreme points of an assignment polytope has been proposed as an implementation tool for solving various types of programming problems. Cabot and Francis [3] have developed a procedure which ranks the extreme points of an assignment polytope in order to generate a monotone increasing sequence of lower bounds for solving concave minimization transportation problems. Fluharty [6] has used similar ideas for solving the quadratic assignment problem. As reported by McKeown [10], this scheme works well principally when the linear term in the objective function numerically dominates the nonlinear term. Murty, Karel and Little [12] have recommended the solution of the traveling salesman problem through the use of ranking extreme points of an assignment polytope. McKeown [8] also suggests the use of such a scheme for solving the bottleneck assignment problem. Sherali [15] has proposed the ranking of possible objective values taken on by a linear assignment problem in the process of solving the quadratic assignment problem. Furthermore, in some applications, a linear assignment problem may need to be solved in the presence of complicating side constraints, some of which may possibly be qualitative such as behavioral constraints [8]. Therefore, the ranking of extreme points of an assignment polytope is a viable approach.

*This research is supported under NSF Grant #ENG77-07468.

As a result, several researchers have devised schemes for ranking the extreme points of an assignment polytope. Murty [11] has proposed a method for the general linear fixed charge problem. This method was implemented, in vain, by Cabot and Francis [3] for the transportation and the assignment polytopes. Murty's procedure is essentially based on the result that given the first, second, ..., kth ranked extreme points of a linear programming problem, the $(k + 1)$st ranked extreme point is geometrically adjacent to at least one of these k extreme points. In the absence of degeneracy, the simplex method may be used to maintain a list of such adjacent extreme points. In the presence of degeneracy, however, one would also need to determine all basic representations of a degenerate vertex in order to use the simplex method. Rubin [14] has shown how Chernikova's algorithm [4,5] may be modified to handle the degenerate case. Further, McKeown and Rubin [9] have specialized this technique for ranking extreme points of a transportation problem, and McKeown [8] has exploited the special structure of the assignment polytope to further specialize this scheme for the linear assignment problem.

Basically, the principal difficulty with these schemes arises from the nature of the assignment polytope itself. Balinski and Russakoff [2] have demonstrated that for an assignment problem of size m, there are $\sum_{i=0}^{m-2} = \binom{m}{i} (m - i - 1)!$ extreme points adjacent to any given extreme point. For $m = 8$, this figure is 16,064. Moreover, the assignment polytope has a diameter of two. That is, given any two extreme points, x and y, either x and y are geometrically adjacent to each other, or there exists an extreme point z distinct from x and y such that x and z are geometrically adjacent and so are y and z. The implication of this is that the procedures of Murty [11], Rubin [14] and McKeown and Rubin [9] which require an explicit listing of adjacent extreme points become intractable for the linear assignment problem. Even the procedure of McKeown [8], which uses cost considerations in order to eliminate the storage of a subset of edges adjacent at an extreme point, is plagued with this problem. This viewpoint is further supported by the storage overflow problems reported by Fluharty [6] and McKeown [10].

This paper proposes a procedure for ranking the extreme points of an assignment polytope which completely avoids the search over adjacent extreme points of the ranked solutions. The technique employs trivially generated cutting planes in an implicit enumeration framework, and is capable of ranking any number of extreme points at any feasible value for the assignment problem. Thus, in particular, it can rank all the extreme points of the assignment polytope, or it can rank the objective function values that can be taken on by the linear assignment problem by simply ranking only one extreme point at each objective value, or it can rank the objective function values that can be taken on by the linear assignment problem by simply ranking only one extreme point at each objective value, or it can rank a preset number of extreme points at each attainable value within any possible range of objective function values. Moreover, the ranking of extreme points at any given objective value is independent of the ranking at other values, and hence storage problems are greatly reduced.

In the next section, we discuss the proposed scheme and then illustrate it through a numerical example. Finally, we present computational results using several test problems including some available in the literature.

2. A TWO-PHASE PROCEDURE FOR RANKING THE EXTREME POINTS OF AN ASSIGNMENT PROBLEM

Consider the linear assignment problem

$$\text{minimize} \left\{ \sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} x_{ij} \colon x \in X_A \right\}$$

where

$$(2.1) \qquad \begin{aligned} X_A = \{x = (x_{11}, \ldots, x_{mm}) \colon & \sum_{i=1}^{m} x_{ij} = 1 \text{ for each } j = 1, \ldots, m, \\ & \sum_{j=1}^{m} x_{ij} = 1 \text{ for each } i = 1, \ldots, m, \\ & x_{ij} = 0 \text{ or } 1 \text{ for each } i, j = 1, \ldots, m\}. \end{aligned}$$

Let $c^t = (c_{11}, \ldots, c_{mm})$, where a superscript t defines the transpose operation, and let

$$(2.2) \qquad \nu_{\min} = \text{minimum } \{c^t x \colon x \in X_A\}, \ \nu_{\max} = \text{maximum } \{c^t x \colon x \in X_A\}.$$

Consider a $\bar{\nu} \in [\nu_{\min}, \nu_{\max}]$ and assume that an $x^1 \in X_A$ is known such at $c^t x^1 = \bar{\nu}$.

In order to demonstrate the capabilities of our procedure as discussed in the foregoing section, we will show how one may rank up to a maximum of $n_{\max}$, say, extreme points having an objective value of $\bar{\nu}$, and then how one may generate another extreme point of the assignment polytope having the smallest objective value greater than $\bar{\nu}$, in case $\nu < \nu_{\max}$. We note that a typical linear assignment problem may have a large number of extreme points with the same objective value $\bar{\nu}$ in the range $[\nu_{min}, \nu_{max}]$. The following scheme which iterates between two phases is designed to exploit this property.

To begin with, in the first phase, denoted by Phase I, all pairwise exchanges of facilities relative to the assignment x^1 which yield the same objective value are attempted. Thus, if $x_{ij}^1 = x_{kl}^1 = 1$ and if $c_{ij} + c_{kl} = c_{il} + c_{kj}$ then an alternative solution with assignments given by x^1 except that the locations of facilities i and k are interchanged, is generated and stored in a list $L(\bar{\nu})$. After all extreme points with objective value $\bar{\nu}$ obtainable through such pairwise exchanges on x^1 have been generated, the next solution x^2, say, is selected from the list $L(\bar{\nu})$, if such a solution exists. Again, alternative solutions having an objective value $\bar{\nu}$ which are obtainable through pairwise exchanges on x^2 are generated and are stored in $L(\bar{\nu})$ provided they have not been generated through pairwise interchanges on previously examined solutions (x^1 in this case). This process is continued till none of the solutions stored in $L(\bar{\nu})$ lead to any new points of X_A having a value of $\bar{\nu}$. At this point, the procedure transfers to the second phase, Phase II. Of course, if at any point in this generation scheme, the number of points in $L(\bar{\nu})$ equals $n_{\max}$, then we may terminate further ranking of solutions having a value $\bar{\nu}$.

In Phase II, an attempt is made to determine if there exists any point in $X_A - L(\bar{\nu})$ which has an objective value of $\bar{\nu}$. Toward this end, consider the solution $x^1 \in X_A$ and define the set

$$(2.3) \qquad I_1 = \{(i, j) \colon x_{ij}^1 = 1\}.$$

Now, it is easy to see that the cutting plane $\sum_{(i,j) \in I_1} x_{ij} \leqslant m - 2$ deletes the point x^1, but no other point in X_A. This is so, because every other point in X_A must have at least two of the variables x_{ij}, $(i, j) \in I_1$, equai to zero. But, if exactly two variables x_{ij}, $(i, j) \in I_1$ are zero in a given solution, then this solution is necessarily obtained from x^1 through a pairwise exchange. Since we are no longer interested in solutions obtained through pairwise exchanges on x^1 in our search for points in X_A of value $\bar{\nu}$, we may impose the cut

$$\sum_{(i,j)\in I_1} x_{ij} \leqslant m-3 \tag{2.4}$$

For subsequent solutions x^2, x^3, ... in the list $L(\bar{\nu})$, a cut of the type (2.4) needs to be generated only if pairwise exchanges on these solutions lead to other solutions in the list $L(\bar{\nu})$. Hence, if the current list contains n points, then we would have some $\bar{n} \leqslant n$ cuts of the type (2.4) generated from solutions x^p for $p \in J$, where J is an appropriate index set with $|J| = \bar{n}$. These cuts would be of the form

$$\left.\begin{array}{l} \displaystyle\sum_{(i,j)\in I_p} x_{ij} \leqslant m-3 \\ \text{where, } I_p = \{(i,j): x_{ij}^p = 1\} \end{array}\right\} \text{ for each } p \in J. \tag{2.5}$$

Note that the cuts (2.5) need not be explicitly generated and stored; it is sufficient to merely maintain the index set J. The reason for this is the following. Consider an enumerated extreme point x^p with $x^p_{i,p(i)} = 1$ for $i = 1, \ldots, m$, where $p(i)$ denotes the location of facility i in the solution x^p. Then one may store x^p in the list $L(\bar{\nu})$ as simply the permutation vector $p(1), \ldots, p(m)$. Now, if $p \in J$, then the cut (2.5) is clearly $\sum_{i=1}^{m} x_{i,p(i)} \leqslant m-3$. Hence, in order to check if a given extreme point x^q with $x^q_{i,q(i)} = 1$ for $i = 1, \ldots, m$ violates or satisfies (2.5), one simply needs to check whether or not the cardinality of the set $\{i \in \{1, \ldots, m\}: p(i) = q(i)\}$ is greater than $m-3$.

Now, in addition to the cuts (2.5), since we are no longer interested in points of objective value less than $\bar{\nu}$, we may impose the cost cut

$$\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}x_{ij} \geqslant \bar{\nu}. \tag{2.6}$$

Let us next define the set

$$Q = \left\{x: \sum_{(i,j)\in I_p} x_{ij} \leqslant m-3 \quad \text{for each } p \in J,\ c'x \geqslant \bar{\nu}\right\} \tag{2.7}$$

and consider the following problem

$$\text{FEAS } (\bar{\nu}):\ \text{minimize } \{c'x: x \in Q \cap X_A\}.$$

Clearly, Problem FEAS $(\bar{\nu})$ will either generate as an optimal solution a point in $X_A - L(\bar{\nu})$ having a value $\bar{\nu}$, or failing which, will indicate that no such point exists. Thus, let $\bar{x}$ solve FEAS $(\bar{\nu})$. If $c'\bar{x} = \bar{\nu}$, then we may set $x^{n+1} = \bar{x}$ and transfer to Phase I, attempting all pairwise exchanges starting with the solution x^{n+1}. On the other hand, if $c'\bar{x} > \bar{\nu}$ or if FEAS $(\bar{\nu})$ is infeasible, then the list $L(\bar{\nu})$ is complete, and we will say that we have completed one *iteration*. Now observe that during the course of an iteration one may have to resort to several solutions of Problem FEAS $(\bar{\nu})$ with different sets Q. As a computational expedient for solving these problems in a particular iteration, we will describe an implicit enumeration scheme which is initialized only once during the first visit to Phase II of the procedure, and is simply updated at each subsequent visit in that iteration.

The fundamental implicit enumeration scheme we use is due to Geoffrion's [7] adaptation of Balas' [1] method. Basically, we chose this depth-first enumeration scheme because of the ease and the minimal storage requirements with which one may perform all the bookkeeping operations necessary in the implicit enumeration procedure. The bookkeeping here involves simply the maintenance of a partial solution list which indicates in chronological order which variables are currently fixed in value at either zero or one. Thus, a partial solution list $(x_{11} = 1, x_{23} = 0, x_{34} = 0, x_{35} = 1)$ would indicate that the variables x_{11}, x_{23}, x_{34} and x_{35} are currently fixed at the indicated values, and moreover, that x_{11} was fixed at one before x_{23} was fixed at zero, which in turn occurred before it was decided to set $x_{34} = 0$, and so on. For the assignment problem, this feature is very convenient for one may enforce the assignment constraints by merely ensuring that no facility is ever assigned to more than one location and vice versa in the partial solution list. In other words, $x_{ij} = 1$ implies $x_{kj} = x_{il} = 0$ for each k, $l = 1, \ldots, m$, $k \neq i$, $l \neq j$. Of course, if a facility is barred from each location, or vice versa, then the partial solution will be fathomed. Thus, for example, if our current partial solution list is $(x_{11} = 1, x_{23} = 0, x_{34} = 0, x_{35} = 1)$, we would never consider augmenting this list by setting $x_{21} = 1$ or $x_{13} = 1$, since $x_{11} = 1$ implies that $x_{21} = 0$ and that $x_{13} = 0$. Furthermore, if this partial solution list is always incremented each time by assigning a facility to a location, then whenever a partial solution is fathomed, one simply locates the rightmost variable in this partial solution list which is fixed at value one, then one complements the value of this variable to zero, and deletes all fixed variables to the right of this variable. Hence, a new partial solution list is obtained. Consequently, consider the partial solution list $(x_{12} = 1, x_{33} = 0, x_{34} = 0, x_{25} = 1, x_{31} = 0)$. Note that since the partial solution list is always augmented by suitably setting some variable equal to one, each of the variables which are currently zero were at some previous point in time equal to one. In particular, x_{31} was equal to one before being set equal to its current value of zero. Observe that the current partial solution list bars facility 3 from all locations 1, 2, ..., 5 (assuming $m = 5$). Thus, this solution is said to be fathomed. Consequently, we locate the rightmost variable in the list which is currently one, namely x_{25}, *complement it to zero, i.e., set $x_{25} = 0$, and delete all variables to its right, namely, eliminate* $x_{31} = 0$ from the list in this case. This gives us the new partial solution list $(x_{12} = 1, x_{33} = 0, x_{34} = 0, x_{25} = 0)$. As proven by Geoffrion [7], this procedure would result in a nonredundant, exhaustive, implicit enumeration of all solutions.

The algorithmic statement below gives the details for the incrementing and the fathoming operations. Before discussing these implementation details, we draw the reader's attention to two specific points. Firstly, the moment a feasible solution to FEAS $(\bar{\nu})$ of value $\bar{\nu}$ is detected, then this solution necessarily solves FEAS $(\bar{\nu})$. Secondly, during the implicit enumeration scheme, we maintain an incumbent solution $\hat{x}$ of value $\hat{\nu}$ which is updated each time an improved solution is found, except if this improved solution has a value $\bar{\nu}$. This ensures that when we have finally enumerated all solutions having a value of $\bar{\nu}$, and have consequently solved the final problem FEAS $(\bar{\nu})$, the incumbent solution value is equal to the next value larger than $\bar{\nu}$, taken on by a feasible assignment, if such a value exists. A detailed description of both Phase I and Phase II of the algorithm is given below. Here, the algorithm finds at most $n_{\max}$ extreme points having a specified objective value equal to $\bar{\nu}$. It is assumed that the cost coefficients c_{ij}'s are all integer valued.

PHASE I. (Generate an initial list $L(\bar{\nu})$ using pairwise exchanges.)

INITIALIZATION: (This step is performed each time a new iteration is being commenced; otherwise, the procedure transfers to Step 1 at each subsequent visit to Phase I.)

Suppose that $x^1 \in X_A$ is given with $c'x^1 = \bar{\nu}$. Let x^1 be the first solution in a list $L(\bar{\nu})$. Set $\hat{\nu} = \infty$, $r = n = 1$, $J = \phi$. Proceed to Step 1.

STEP 1: Attempt all pairwise exchanges on x^r to find solutions of value $\bar{\nu}$ which have not as yet been listed in $L(\bar{\nu})$. Store these solutions by adding them to $L(\bar{\nu})$. Further if either this is the first time Step 1 is being executed at the current visit to Phase I or if pairwise exchanges on x^r lead to new alternative solutions of value $\bar{\nu}$, replace J by $J \cup \{r\}$. In any case, increment n, if necessary, so that n equals the total number of extreme points in the list $L(\bar{\nu})$. If $n \geqslant n_{\max}$, the maximum number of points desired to be ranked at the value $\bar{\nu}$, then transfer to Step 10 of Phase II. Otherwise, continue.

STEP 2: If r is equal to n, then proceed to Phase II. Otherwise, increment r by one and return to Step 1.

PHASE II. (Solution of the current Problem FEAS $(\bar{\nu})$.)

INITIALIZATION: (This step is executed only on the first visit of the procedure to Phase II during a particular iteration. At all subsequent visits, the procedure may transfer directly to Step 3 since the previous partial solution list at the last visit to Phase II is also valid at this stage. The reason being that the new problem FEAS $(\bar{\nu})$ is a further restriction of the previous FEAS $(\bar{\nu})$, and the manner in which Steps 3 through 9 below are executed ensures that solutions previously fathomed are also currently inconsequential.)

Initialize the partial solution vector by assigning $m - 2$ facilities as in the solution x^1. Set $\hat{\nu} = \infty$. Since neither of the two completions of this partial solution lead to a new solution of value $\bar{\nu}$, proceed to Step 6.

STEP 3: Find the feasible completion $\bar{x} \in X_A$ of the partial solution vector which minimizes the value of $c'x$ through the solution of the associated linear assignment problem. That is. solve the linear assignment problem of minimizing $c'x$ subject to $x \in X_A$ given that the assignments indicated by the partial solution list are forced a priori. (Note that by virtue of Step 7 below, this step is never entered with fewer than three unassigned facilities.) If no completion feasible to the assignment constraints exists, or if the resulting solution $\bar{x}$ satisfies $c'\bar{x} \geqslant \hat{\nu}$, transfer to Step 6. Otherwise, continue.

STEP 4: Determine if $\bar{x} \in Q$, where Q is defined by Equation (2.7). Transfer to Step 7 if it is not and continue otherwise.

STEP 5: If $c'\bar{x} = \bar{\nu}$, then set $x^{r+1} = \bar{x}$, replace r by $r + 1$, place $\bar{x}$ in $L(\bar{\nu})$ and go to Step 1 of Phase I. The current problem FEAS $(\bar{\nu})$ is solved, with the optimal solution value being $\bar{\nu}$. Otherwise, since $c'\bar{x} < \hat{\nu}$ from Step 3, set $\hat{x} = \bar{x}$ and $\hat{\nu} = c'\bar{x}$ and continue.

STEP 6: Fathom the current partial solution by finding the rightmost variable in the partial solution list which is fixed at one, complementing it to zero, and deleting (setting free) all fixed variables to the right of this variable. If no such variable exists, go to Step 10 since then the current problem FEAS $(\bar{\nu})$ is solved with the optimal objective function value being $\hat{\nu} > \bar{\nu}$. Otherwise, return to Step 3.

STEP 7: If more than three facilities have not as yet been located somewhere in the current partial solution, proceed to Step 8. Otherwise, enumerate the feasible completions in X_A of the partial solution. Note that at most six such completions exist. If either there are no feasible completions or if all the feasible completions are infeasible to Q, then transfer to Step 6.

If any feasible completion lies in Q and has an objective value $\bar{\nu}$, then set x^{r+1} equal to this solution, place x^{r+1} in the $L(\bar{\nu})$, replace r by $r + 1$ and transfer to Phase I.

If any feasible completion lies in Q and has a value less than $\hat{\nu}$, then pick that solution with the least objective value and update $\hat{x}$ and $\hat{\nu}$ by replacing them with this solution and its objective value respectively and transfer to Step 6. If none of the above cases hold, simply transfer to Step 6.

STEP 8: Find a feasible completion $\bar{\bar{x}} \in X_A$ which maximizes the value of $c'x$. If $c'\bar{\bar{x}} < \bar{\nu}$, then go to Step 6 since then clearly there exists no completion of the current partial solution feasible to FEAS $(\bar{\nu})$. Otherwise, continue.

STEP 9: Let $\bar{x} \in X_A$ be the solution found in Step 3. Increment the current partial solution by assigning that unassigned facility i to that free location j which satisfies $\bar{x}_{ij} = 1$ and which yields the smallest value of c_{ij} from among all such pairs (i, j). If there are more than three unassigned facilities, go to Step 3. Otherwise, go to Step 7.

STEP 10: The list $L(\bar{\nu})$ is complete. If $\bar{\nu} = \nu_{\max}$, the procedure terminates. If it is now required to generate a new point $x^* \in X_A$ with $c'x^* = \nu^*$ such that ν^* is the next greater value taken on by the linear assignment objective function after $\bar{\nu}$, perform the following routine.

First check if $\hat{\nu} = \bar{\nu} + 1$. If it is, then necessarily, $x^* = \hat{x}$, $\nu^* = \hat{\nu}$. If not, then using the current $\hat{x}$ as an incumbent solution of value $\hat{\nu}$ and setting $Q = \{x: c'x \geqslant \bar{\nu} + 1\}$, use the above implicit enumeration scheme (Phase II) with obvious modifications to solve the problem of minimizing $c'x$ subject to $x \in X_A \cap Q$. Then x^* and ν^* are the optimal solution and optimal objective value obtained, respectively. Again, note that if during this search, a solution of value $\bar{\nu} + 1$ is detected, then this is an optimal solution. Now, replacing $\bar{\nu}$ by ν^* and x^1 by x^*, one may initiate a new list $L(\bar{\nu})$ with x^1 as the first solution in this list and transfer to Phase I.

Observe again that during the performance of Phase II if any solutions are fathomed, then we will never have the occasion to reexamine these solutions during subsequent visits to Phase II for a particular iteration. Hence, the implicit enumeration scheme of Phase II may simply be updated at each visit. For this reason, the incumbent solution $\hat{x}$ is not updated when Phase II detects an optimal solution to FEAS $(\bar{\nu})$ of value $\bar{\nu}$.

This further enables one to possibly avoid extra search in determining the solution x^* in Step 10. In our experience, the condition $\hat{\nu} = \bar{\nu} + 1$ often held at Step 10. Finally, we explain the motivation behind Step 7. Observe that the cuts of type (2.5) are such that if at least three facilities are unassigned in a partial solution then for each of these cuts, there exists a completion (not necessarily feasible, though) which will satisfy that cut. It is because of this reason that we do not attempt to check for existence of feasible completions which satisfy these cuts, until three facilities remain unassigned. However, we do check at Step 8 for existence of feasible completions which satisfy the cut $c'x \geqslant \bar{\nu}$.

3. ILLUSTRATIVE EXAMPLE

Consider a linear assignment problem with $m = 4$, having the following cost matrix

$$[c_{ij}] = \begin{bmatrix} 3 & 2 & 1 & 3 \\ 1 & 2 & 3 & 4 \\ 3 & 1 & 2 & 4 \\ 1 & 2 & 2 & 1 \end{bmatrix}.$$

For this problem, $\nu_{\min} = 4$, $\nu_{\max} = 12$. Let us select $\bar{\nu} = \nu_{\min} = 4$ and enumerate all possible solutions with this objective value. Hence, we know $x^1 = (3,1,2,4)$ of value $c'x^1 = \bar{\nu} \equiv 4$. (Note that an assignment solution is being represented as a permutation vector $(p(1), P(2), \ldots, p(m))$, where $p(i)$ satisfies $x_{ip(i)} = 1$, $i = 1, \ldots, m$ for that solution.)

ITERATION 1.

PHASE I.

INITIALIZATION: Given $x^1 = (3,1,2,4)$ with $c'x^1 = \bar{\nu} = 4$, we have, $L(\bar{\nu}) = \{(3,1,2,4)\}$. Set $\hat{\nu} = \infty$, $r = n = 1$, $J = \phi$.

STEP 1: No further solutions are generated. Set $J = \{1\}$, $\bar{n} = |J| = 1$ and from Equation (2.7), $Q = \{x: x_{13} + x_{21} + x_{32} + x_{44} \leqslant 1,\ c'x \geqslant 4\}$.

STEP 2: Since $r = n = 1$, proceed to Phase II.

PHASE II.

INITIALIZATION: Here, $m - 2 = 2$. Thus, our starting partial solution list, abbreviated by PS henceforth, is $PS = (x_{13} = 1,\ x_{21} = 1)$.

STEP 6: Fathom PS. This gives the updated $PS = (x_{13} = 1,\ x_{21} = 0)$.

STEP 3: Solve the assignment problem with $x_{13} = 1$, $x_{21} = 0$ to get $\bar{x} = (3,4,2,1)$ with $c'\bar{x} = 7 < \hat{\nu} = \infty$.

STEP 4: $\bar{x} \notin Q$.

STEP 7: Currently, $PS = (x_{13} = 1,\ x_{21} = 0)$. Thus, the completions are
(3,2,1,4) $\notin Q$
(3,2,4,1) $\in Q$ of value 8
(3,4,1,2) $\in Q$ of value 10
(3,4,2,1) $\notin Q$.

Hence, update $\hat{x} = (3,2,4,1)$ with $\hat{\nu} = 8$.

STEP 6: Fathom PS. The new $PS = (x_{13} = 0)$.

STEP 3: Solve the assignment problem with $x_{13} = 0$ to get $\bar{x} = (2,1,3,4)$ with $c'\bar{x} = 6 < \hat{\nu} = 8$.

STEP 4: $\bar{x} \notin Q$

STEP 7: Currently, $PS = (x_{13} = 0)$. Thus, all four facilities are unassigned.

STEP 8: Putting $x_{13} = 0$, find the maximizing assignment solution $\bar{\bar{x}}$. Thus, $c'\bar{\bar{x}} = 12 \not< \bar{\nu} = 4$.

STEP 9: Examine $\bar{x} = (2,1,3,4)$ found in Step 3 above. Force the assignment $x_{21} = 1$. Hence, the new $PS = (x_{13} = 0, x_{21} = 1)$.

STEP 7: Enumerate the completions of PS
(2,1,3,4) $\notin Q$
(2,1,4,3) $\in Q$ of value 9
(4,1,2,3) $\notin Q$
(4,1,3,2) $\in Q$ value 8

STEP 6: Fathom PS. The new $PS = (x_{13} = 0, x_{21} = 0)$.

STEP 3: Solve the assignment problem with $x_{13} = 0$, $x_{21} = 0$ to get $\bar{x} = (1,2,3,4)$ with $c'\bar{x} = 8 \geqslant \hat{\nu} = 8$.

STEP 6: Since $PS = (x_{13} = 0, x_{21} = 0)$ currently, fathoming PS leads to a termination of the present iteration.

STEP 10: $L(4) = \{(3,1,2,4)\}$, a singleton. Now, to find $x^* \in X_A$ of Step 10, since $\hat{\nu} = 8 > \bar{\nu} + 1 = 5$, we let $Q = \{x: c'x \geqslant 5\}$ and using $\hat{x} = (3,2,4,1)$, we construct a starting partial solution list $PS = (x_{13} = 1, x_{22} = 1)$. Verifying that the best completion of this list is indeed $\hat{x}$ of value $\hat{\nu} = 8$, we proceed to Step 6.

STEP 6: Fathom PS. The new $PS = (x_{13} = 1, x_{22} = 0)$.

STEP 3: With $x_{13} = 1$, $x_{22} = 0$, the linear assignment problem yields $\bar{x} = (3,4,2,1)$ with $c'\bar{x} = 7 < \hat{\nu} = 8$.

STEP 4: $\bar{x} \in Q$ of Step 10 above.

STEP 5: Update $\hat{x} \equiv \bar{x} = (3,4,2,1)$, $\hat{\nu} = c'\bar{x} = 7$.

STEP 6: Fathom PS. The new $PS = (x_{13} = 0)$.

STEP 3: With $x_{13} = 0$, solve the linear assignment problem to get $\bar{x} = (2,1,3,4)$ with $c'\bar{x} = 6 < \hat{\nu} = 7$.

STEP 4: $\bar{x} \in Q$ of Step 10 above.

STEP 5: Update $\hat{x} \equiv \bar{x} = (2,1,3,4)$, $\hat{\nu} = c'\bar{x} = 6$.

STEP 6: Fathom PS. [illegible]ds to a termination of the operation of finding the next larger objective value. Hence, [illegible] ν^* of Step 10 are $x^* \equiv \hat{x} = (2,1,3,4)$ and $\nu^* \equiv \hat{\nu} = 6$.

One may now enumerate solutions of value 6 if so desired.

4. COMPUTATIONAL EXPERIENCE

For the purpose of computational testing, we will use the following test problems. The problem of size $m = 4$ is the illustrative example of Section 3. The problem of size $m = 3$ has the following cost matrix

$$\begin{bmatrix} 3 & 2 & 1 \\ 1 & 2 & 1 \\ 2 & 1 & 2 \end{bmatrix}$$

Problems of sizes $m = 5, 6, 7, 8, 12$ and 15 have as their cost matrices those corresponding to the linear assignment problems, the extreme points of which need to be ranked by Cabot and Francis' [3] procedure applied to the quadratic assignment test problems of Nugent, Vollmann and Ruml [13] of respective sizes $m = 5, 6. 7, 8, 12$ and 15. Problems of sizes 25, 50, 75, and 100 have their cost data generated as indicated below.

The procedure was coded in FORTRAN IV and the computational times reported are in c.p.u. seconds of execution time on a CDC Cyber 70 Model 74-28/CDC 6400 computer in time-sharing mode under normal batch operation with compilation performed using the OPT = 1 option. Table 4.1 gives the experience for a total ranking of all extreme points of the assignment polytope, that is, with $n_{\max} = m!$. Table 4.2 gives the experience for simply ranking the objective function values taken on by the linear assignment problem, that is, with $n_{\max} = 1$. Table 4.3 gives results for ranking at most 500 extreme points at each objective value assumed by the linear assignment problem, so that $n_{\max} = 500$. Table 4.4 deals with ranking at most 50 extreme points at the first 10 objective function values attainable. The data for these problems is generated according to $c_{ij} = (ij) \mathrm{mod}\ 10$. Finally, Table 4.5 gives some computational testing of problems with cost coefficients uniformly generated at random in the interval [0, 100].

Note that during any iteration, each visit to Phase II results in only one additional extreme point being enumerated except for the last visit, when no extreme point is enumerated. In the tables below, the column designated "a" records the average number of passes through Phases I and II per iteration. Hence, the total number of extreme points generated in Phase II is given by (a-1) (Number of iterations).

TABLE 4.1 — *Ranking of All Extreme Points* ($n_{\max} = m!$)

m	$\nu_{\min}$	$\nu_{\max}$	Number of Iterations	a	b	c.p.u. Seconds of Execution Time
3	3	7	3	1.00	1(3), 4(5), 1(7)	0.018
4	4	12	8	2.25	1(4), 1(6), 3(7), 7(8), 3(9), 4(10), 4(11), 1(12)	0.477
5	50	55	6	2.833	20(50), 14(51), 28(52), 28(53), 16(54), 14(55)	1.378
6	82	84	2	1.00	480(82), 240(84)	66.677
7	137	144	8	1.125	48(137), 384(138), 912(139), 1344(140), 1200(141), 768(142), 288(143), 96(144),	981.556

a — Average number of passes through Phases I and II per iteration

b — Number of extreme points ranked with values $\bar{\nu}$ in parentheses at each iteration

TABLE 4.2 — *Ranking of All Objective Function Values* ($n_{\max} = 1$)

m	$\nu_{\min}$	$\nu_{\max}$	Number of Iterations	a	Execution Time in c.p.u Seconds		
					min time per iteration	max time per iteration	Total time
3	3	7	3	1.00	0.002	0.004	0.008
4	4	12	8	1.00	~0.000	0.058	0.092
5	50	55	6	1.00	~0.000	0.026	0.042
6	82	84	2	1.00	0.002	0.101	0.103
7	137	144	8	1.00	~0.000	0.196	0.443
8	186	199	14	1.00	~0.000	1.723	5.670
12	493	517	25	1.00	~0.000	1.622	5.700
15	963	1034	72	1.00	~0.000	4.320	26.631

a — Same connotation as for Table 4.1.

TABLE 4.3 — *Partial Ranking of Extreme Points at each Objective Value with* $n_{\max} = 500$

m	$\nu_{\min}$	$\nu_{\max}$	Number of Iterations	a	c.p.u. Seconds of Execution Time
7	137	144	8	1.125	166.142
8	186	199	14	1.00	250.906
12	493	517	25	1.00	201.315
15	963	1034	72	1.00	624.209

a — Same connotation as for Table 4.1

TABLE 4.4 — *Partial Ranking of Extreme Points at the First 10 Objective Values with* $n_{\max} = 50$

m	$\nu_{\min}$	$\nu_{\max}$	a	Execution Time in c.p.u Seconds		
				min time per iteration	max time per iteration	Total time
25	21	183	1.00	0.034	0.288	0.900
50	40	365	1.00	0.026	0.102	0.522
75	61	548	1.00	0.032	0.354	1.046
100	80	730	1.00	0.042	0.154	0.828

a — Same connotation as for Table 4.1.

TABLE 4.5 — *Partial Ranking of Extreme Points of Problems with Cost Coefficients Uniformly Generated at Random in the Interval [0, 100]*

Number of Iterations and $n_{\max}$	m	$\nu_{\min}$	$\nu_{\max}$	a	Execution Time in c.p.u Seconds		
					min time per iteration	max time per iteration	Total time
Number of Iterations = 10 $n_{\max} = 1$	25	99	2356	1.00	~0.00	22.46	46.47
	50	107	4805	1.00	~0.00	15.91	33.29
	75	108	7297	1.00	~0.00	15.25	33.15
	100	135	9794	1.00	~0.00	51.27	105.19
Number of Iterations = 5 $n_{\max} = 5$	25	99	2356	3.00	2.63	23.59	66.20
	50	107	4805	4.20	24.55	98.21	273.23
	75	108	7297	4.00	81.68	184.47	585.87
	100	135	9794	3.00	60.38	171.23	609.47
Number of Iterations = 10 $n_{\max} = m!$	25	99	2356	6.00	2.25	56.06	282.07
Number of Iterations = 5 $n_{\max} = m!$	25	99	2356	3.20	2.22	25.95	70.03

a — *Same connotation as for Table 4.1*

REFERENCES

[1] Balas, E., "An Additive Algorithm for Solving Linear Program with Zero-One Variables," Operations Research, *13*, 517-546 (1965).

[2] Balinski, M.L. and A. Russakoff, "On the Assignment Polytope," SIAM Review, *16*, 516-525 (1974).

[3] Cabot, A.V. and R.L. Francis, "Solving Certain Nonconvex Quadratic Minimization Problems by Ranking the Extreme Points," Operations Research, *18*, 82-86 (1970).

[4] Chernikova, N.V., "Algorithm for Finding a General Formula for the Nonnegative Solutions of a System of Linear Equations," U.S.S.R. Computational Mathematics and Mathematical Physics, *4*, 151-158 (1964).

[5] Chernikova, N.V., "Algorithm for Finding a General Formula for the Nonnegative Solutions of a System of Linear Inequalities," U.S.S.R. Computational Mathematics and Mathematical Physics, *5*, 228-233 (1965).

[6] Fluharty, R., "Solving Quadratic Assignment Problems by Ranking the Assignments," unpublished Master's Thesis, Ohio State University, Columbus, OH (1970).

[7] Geoffrion, A.M., "An Improved Implicit Enumeration Approach for Integer Programming," Operational Research, *17*, 437-454 (1969).

[8] McKeown, P.G., "Determining Adjacent Vertices on Assignment Polytopes," Naval Research Logistics Quarterly, *23*, 455-460 (1976).

[9] McKeown, P.G. and D.S. Rubin, "Adjacent Vertices on Transportation Polytopes," Naval Research Logistics Quarterly, *22*, 365-374 (1975).

[10] McKeown, P.G., "Extreme Point Ranking Algorithms: A Computational Survey," Computers and Mathematical Programming, NBS Special Publication, *502*, 216-222 (1978).

[11] Murty, K.G., "Solving the Fixed Charge Problem by Ranking the Extreme Points," Operations Research, *16*, 268-279 (1968).

[12] Murty, K.G., C. Karel and J.D.C. Little, "The Travelling Salesman Problem: Solution by a Method of Ranking Assignments," unpublished report, Case Institute of Technology, Cleveland, OH (1962).

[13] Nugent, C.E., T.E. Vollmann and J. Ruml, "An Experimental Comparison of Techniques for the Assignment of Facilities to Locations," Operations Research, *16*, 150-173 (1968).

[14] Rubin, D.S., "Neighboring Vertices on Convex Polyhedral Sets," unpublished report, University of North Carolina, Chapel Hill, NC (Aug. 1972).

[15] Sherali, H.D., "The Quadratic Assignment Problem: Exact and Heuristic Methods," unpublished doctoral dissertation, Georgia Institute of Technology, Atlanta, GA (1979).

A PRIMAL-DUAL CUTTING-PLANE ALGORITHM FOR ALL-INTEGER PROGRAMMING

Parviz Ghandforoush and Larry M. Austin

College of Business Administration
Texas Tech University
Lubbock, Texas

ABSTRACT

The integer programming literature contains many algorithms for solving all-integer programming problems but, in general, existing algorithms are less than satisfactory even in solving problems of modest size. In this paper we present a new technique for solving the all-integer, integer programming problem. This algorithm is a hybrid (i.e., primal-dual) cutting-plane method which alternates between a primal-feasible stage related to Young's simplified primal algorithm, and a dual-infeasible stage related to Gomory's dual all-integer algorithm. We present the results of computational testing.

In this paper, we describe an algorithm which we call the Constructive Primal-Dual Algorithm (CPDA), for solving IP problems of the form (P) below.

$$(1) \qquad \text{maximize} \quad x_0 - \sum_{j=1}^{n} c_j x_j = 0$$

(P)

$$(2) \qquad \text{subject to} \quad \sum_{j=1}^{n} a_{ij} x_j \leqslant b_i, \quad i = 1, \ldots, m$$

$$(3) \qquad x_j \geqslant 0 \text{ and integer}, \; j = 1, \ldots, n$$

$$(4) \qquad \text{where} \quad c_j, \; a_{ij}, \; b_i \text{ are integer for all } i, j.$$

Glover [2] developed a "Pseudo Primal-Dual" integer programming algorithm (PPDA) in which the pure integer programming problem is solved in two stages, systematically violating and restoring dual feasibility while maintaining an all-integer tableau (PPDA is the only primal-dual algorithm reported in the literature). This algorithm starts with a dual feasible tableau similär to that of Gomory [3]. A cut constraint is generated from the row with the largest number of negative elements, and dual simplex is then applied. If this stage does not destroy dual feasibility, then the process is repeated until an optimal solution is reached. Otherwise, dual feasibility is restored by a sequence of "pseudo-primal" pivot steps using the column that is lexicographically most negative when divided by the corresponding coefficient in the source row (restricting attention to positive coefficients), and the pivot row from the dual stage.

The CPDA alternates between a primal-feasible stage related to Young's SPA [7], and a primal-and dual-infeasible stage related to Gomory's dual all-integer algorithm [3]. The CPDA departs from Young's technique in its choice of the cut row. This algorithm solves (P) by usually starting primal-feasible and dual-infeasible (if (P) is a maximization problem with all "$\leqslant$" constraints), as does Young's SPA. Whenever a stationary cycle is encountered, the CPDA avoids degenerate iterations by developing a cut which deliberately moves into the infeasible region, and then attempts to return to primal feasibility at a better solution point than the one from which it departed.

1. THE CONSTRUCTIVE PRIMAL-DUAL ALGORITHM

This section presents an outline and a discussion of the CPDA. Two similar versions of the algorithm are presented (CPDA-1 and CPDA-2), and are used in solving two distinct classes of integer programming problems. We employ the Beale tableau, explicitly carrying only the nonbasic columns and the constraint vector.

CPDA-1 can be used in solving most standard IP problems; i.e., those IP problems that are not highly primal- or dual-degenerate. Algorithmic steps of the CPDA-1 are as follows.

Primal Stage

STEP 1: Write the IP in form (P). If some $b_i < 0$ then go to the dual stage. Otherwise, check to determine whether $c_j \geqslant 0$ for $j = 1, 2, \ldots, n$; if so, stop; the current basis is optimal; otherwise, go to Step 2.

STEP 2: Select the column α_k which is lexicographically the most negative as the pivot column. If all components of α_k are negative, stop; the value of the objective function is unbounded. Otherwise, select as the source row that β_r for which b_i/a_{ik} is a minimum, and $a_{ik} > 0$. Break ties by arbitrary selection. If $b_r \geqslant a_{rk}$ go to Step 4; otherwise, go to Step 3.

STEP 3: In column α_k search for a new row β_r such that $[b_r/a_{rk}]$ is the smallest ratio greater than or equal to 1, where $a_{rk} > 0$, and where [ ] signifies the greatest integer part. If no such row β_r exists then row β_r from Step 2 is selected as the source row. Go to Step 4.

STEP 4: Construct the cut-constraint as follows:

$$\sum_{j=1}^{n} [a_{rj}/a_{rk}]\,(t_j) + s_c = [b_r/a_{rk}], \tag{5}$$

where s_c is a nonnegative integer variable called the "cut-slack." Append constraint (5) to the bottom of the current tableau. Go to Step 5.

STEP 5: Perform a simplex iteration by pivoting on the "+1" (the pivot element) in the bottom row and in the selected pivot column α_k. The slack variable s_c in (5) becomes a nonbasic variable. After the iteration is completed, discard the bottom row. If some $b_i < 0$ in the current tableau, then go to the dual stage. If all $c_j \geqslant 0$, stop; the current basis is optimal. Otherwise, go to Step 2.

Dual Stage

STEP 6: If all $b_i \geqslant 0$ and $c_j \geqslant 0$, stop; an optimal solution has been obtained. Otherwise, go to Step 7.

STEP 7: Select the source-row β_r as the row for which $r \geqslant 1$, and $r = \min\{i: b_i < 0\}$. Go to Step 8.

STEP 8: If $a_{rj} \geqslant 0$ for all $j \geqslant 2$, stop; (P) has no feasible solution. Otherwise, consider all columns α_j such that $a_{rj} < 0$. Choose α_k such that $c_k < 0$ and $\alpha_k \overset{l}{<} a_j$ for all $j = 1, 2, \ldots, n$, where "$\overset{l}{<}$" represents the lexicographically smallest column, and

$$\alpha_j = \left\{\frac{a_{1j}}{a_{rj}}, \frac{a_{2j}}{a_{rj}}, \ldots, \frac{a_{mj}}{a_{rj}}\right\}.$$

If there is no such $c_k < 0$ then go to Step 9; otherwise go to Step 10.

STEP 9: For the source-row β_r from Step 7 and for all $a_{rj} < 0$, select the entering column α_k such that $\alpha_k \overset{l}{>} \alpha_j$ if and only if $c_k > 0$. If there is no $c_j > 0$, then consider α_j such that $c_j = 0$, for $a_{rj} > 0$. Go to Step 10.

STEP 10: Construct a dual cut with cut-constant $|a_{rk}|$, generating a cut with a pivot element of "-1." Append the constraint (5) to the bottom of the tableau. Go to Step 11.

STEP 11: Perform a simplex iteration by pivoting on the "-1" in the bottom row and in the selected pivot column α_k. The slack variable s_c in (5) becomes a nonbasic variable. After the iteration is completed, discard the bottom row. If some $b_i < 0$, then go to Step 6; otherwise go to the primal stage (Step 2).

We modified CPDA-1 to solve a special class of IP problems which are highly primal- and dual-degenerate (e.g., fixed-charge IP problems). A modified version of CPDA-1 (designated herein as CPDA-2) involves the following changes to the original algorithm.

In Step 2 of CPDA-1, if $b_r \geqslant a_{rk}$ for some source-row β_r then go to Step 4. Otherwise, we substitute a procedure for Step 3 which constructs a cut from the objective function of the current tableau, using the largest (least negative) negative c_j as the "cut constant." Inequality (6) is used to construct the cut-constraint (7); that is:

$$\text{(6)} \qquad \sum_{j=1}^{n} c_j x_j \geqslant x_0 + 1,$$

$$\text{(7)} \qquad \sum_{j=1}^{n} [c_j/c_k]\,(t_j) + s_c = [(x_0 + 1)/c_k],$$

where x_0 is the current value of the objective function. Since the resulting tableau is primal-infeasible, we proceed to Step 6.

Note that all the other steps (with the exception of Step 3) remain the same in this version of CPDA.

Discussion of CPDA

At this point, it is useful to discuss in more detail some of the algorithmic steps presented in the previous section.

In Step 2 of CPDA-1, if $b_r/a_{rk} \geqslant 1$ then the next iteration will be a transition cycle; otherwise a normal primal iteration would generate a stationary iteration (since $a_{rk} > b_r$). By imposing the "next higher integer rule," the CDPA is actually attempting to avoid a stationary cycle whenever possible by transforming the primal-feasible problem into a primal- and dual-infeasible one. The "next higher integer" rule is actually the essence of the basic CDPA algorithm. It is at this stage of the algorithm that CPDA departs from Young's SPA.

In Step 8, the effect is to generate the strongest possible cut in the primal-dual-infeasible tableau. Experience with many test problems has shown that primal feasibility may be reached much faster if the c_k/a_{rk} ratio is the smallest possible ratio for all c_k and $a_{rk} > 0$. In most non-degenerate problems (in a dual sense), this rule is very effective. However, for highly dual-degenerate problems, CPDA-1 is usually less efficient. In Step 9, the CPDA is attempting to avoid dual-degenerate iterations whenever possible. That is, when zero prices are present, CPDA refuses to use those columns as pivot elements; instead it searches for a $c_j > 0$ combination.

The major difference between the CPDA-1 and CPDA-2 is in the use of the "Objective cut" rule. That is, instead of selecting the next higher integer ratio of b_r/a_{rk} (where $b_r > a_{rk}$), whenever the next iteration would normally be a stationary cycle, a cut constructed from the current objective function is appended to the bottom of the tableau, and the RHS value replaced by a "-1". The effect is that CPDA-2 is constraining the current value of the objective function to be at least one unit higher, and challenging the primal-feasibility state of the problem. Two conditions may result: (a) the tableau will remain primal-feasible, and generate further transition cycles, increasing the objective function value; or (b) the algorithm will enter a primal-dual-infeasible stage and attempt to regain primal-feasibility status by generating dual cuts. At this point the tableau may indicate that there is no feasible solution to the problem, by recognizing that in at least one of the constraints, $b_i < 0$ and all $a_{ij} \geqslant 0$. In that case, the computation is terminated with the most recent primal-feasible tableau being the optimal solution of the IP problem. However, if this condition does not occur, the CPDA continues with the primal-dual-infeasible stage of the problem until a primal-feasible tableau is reached. This process is repeated until an optimal solution is obtained.

2. RESULTS OF COMPUTATIONAL TESTING

A set of nineteen test problems reported by Haldi [4], and solutions given by Trauth and Woolsey [5] and Wahi and Bradley [6], are used for computational testing. These problems are categorized by Trauth and Woolsey as follows:

(a) ten fixed-charge problems of Haldi; and

(b) nine allocation problems of Trauth and Woolsey.

Trauth and Woolsey report that each problem set has been selected to show a different facet of the generalized integer linear programming problem. The problems are, as a rule, rather small; however, their difficulty is illustrated by the fact that three of the smallest did not converge with four of five commercial codes after 15,000 iterations.

The nine allocation problems are used to investigate the sensitivity of some integer linear problems to a relatively minor change in the problem matrix. Each allocation problem consists of ten decision variables, one major constraint, and the upper bound constraints for each of the decision variables. All nine problems are identical in their structure except for the right-hand-side values. The ten decision variables are of the 0-1 type.

The ten fixed-charge problems were chosen because of the difficulty of solution in spite of their small size. These problems are difficult chiefly because their initial tableaus are heavily primal- and dual-degenerate.

Analysis of Computational Results

Results of computational testing for ten "fixed-charge" problems are exhibited in Table 1. The "fixed-charge" problems are interesting and unusually difficult to solve, because their starting tableaus are heavily dual- and primal-degenerate. The performance of CPDA-2 is compared to that of seven other algorithms. Incidently, the authors' version of Young's SPA did not converge for any of the twenty-five test problems; it is therefore excluded from further analysis in the following "computational results" tables.

TABLE 1 — *Results of Computational Testing for Fixed-Charge Problems*

	On-Site Tests			Commercial Codes[1]				
Code Problems #	CPDA-2 Optimality	Feasibility	BDA	IPM3	LIP1	IL2-1	ILP2-2	IPSC
1	3	7*	12	54	24	135	36	32
2	3	7*	14	81	15	94	47	45
3	3	7*	10	37	26	154	104	56
4	6	8*	12	91	18	93	18	22
5	7	8*	107	+7000	158	+7000	+7000	6104
6	7	8*	83	+7000	123	+7000	311	3320
7	7	8*	107	+7000	159	+7000	+7000	+7000
8	7	8*	83	+7000	126	+7000	306	+7000
9	64	65	44	118	42*	+7000	289	339
10	4**	NA	879	1396	102*	+7000	+7000	+7000

[1] Source: [5]

*Best number of iterations

**Best useful solution (within 20% of optimum)—no convergence in 10 minutes of CPU time

For every fixed-charge problem under CPDA-2 in Table 1, two different results are reported: (a) the number of iterations after CPDA-2 has attained optimality; and (b) the number of iterations after CPDA-2 has reached an infeasible state. For example, in fixed-charge problem #1 CPDA-2 reached the optimum solution in only three iterations. Then, in order to determine whether the current solution is feasible as well, it proceeded with the algorithmic process until, in the seventh iteration, it reached an infeasible state. Therefore, in this example, only three iterations were required by CPDA to solve problem #1, but seven were required to verify that fact.

For problems #1 through #8, CPDA-2 converged in less than eight iterations, which is a remarkably consistent and encouraging performance. It is worth noting that CPDA-2 reached the "near-optimum" solution for these problems almost instantly. For fixed-charge problem #9, CPDA-2 was only outperformed by Austin's Bounded Descent Algorithm [1] and the commercial code LIP1, although the difference in number of iterations does not appear to be large. For the fixed-charge problems, from an overall point of view, CPDA-2 reduced the number of iterations significantly. Next to CPDA-2, Austin's BDA showed an interesting superiority over all the other IP codes examined.

The results of computational testing for the nine allocation problems are exhibited in Table 2. Since the allocation problems are "normal" integer programming problems (i.e., they are not heavily primal- and dual-degenerate), the usefulness of CPDA-2 is limited. Therefore, the basic form of CPDA, i.e., CPDA-1 was used in solving these problems. CPDA-1 was slightly outperformed by BDA and one of the commercial IP codes (IPM3). However, in general, the number of iterations appears to have the same order of magnitude for these algorithms. One reason for this poorer performance by CPDA-1 is that is it a primal-dual algorithm. That is, problems are solved in two distinct stages in which the algorithm searches for both optimality and feasibility conditions, thus requiring a relatively larger number of iterations in reaching the final solution. However, this difficulty did not deter the CPDA from outperforming several of the "one stage" IP codes tested. For instance, in almost all the cases, the performance of CPDA-1 can be ranked no worse than three out of a possible seven (in fact for problems #7 and #8 CPDA gave the best result).

TABLE 2 — *Results of Computational Testing for Allocation Problems*

	On-Site Tests		Commercial Codes[1]				
Code Problems #	CPDA-1	BDA	IPM3	LIP1	IL2-1	ILP2-2	IPSC
1	39	13*	14	19	54	51	46
2	56	20*	31	55	163	77	64
3	23	14*	30	41	168	59	71
4	30	12*	18	19	192	48	62
5	30	3*	11	12	139	32	50
6	45	35	18*	40	157	54	81
7	38*	395	61	81	504	119	131
8	21*	38	21*	51	370	57	102
9	66	2*	12	12	201	34	44

[1] Source: [5]

*Best number of iterations

3. CONVERGENCE PROPERTIES

As regards the existence of a mathematical proof of convergence for the CPDA, two distinct stages must be considered—primal-feasible convergence and primal-dual-infeasible convergence. For the first case, two conditions may occur: (a) the next iteration in the solution process is a transition cycle, in which case the convergence proof will be similar to that for Young's SPA [7]; or (b) the next iteration in the solution process is a stationary cycle, in which case the CPDA will require the use of the "next higher integer rule," which would in fact necessitate a convergence proof of the CPDA for the first stage. No formal mathematical proof has been obtained for this condition. Fortunately, the second stage of CPDA (i.e., the primal-dual-infeasible case) is very similar to that of Glover's PPDA [2] for which a convergence proof is available in the literature.

On the surface, there would seem to be the possibility that the CPDA might leave the feasible region and reenter at the same point from which it departed—thus cycling and failing to converge. However, during dual iterations, the feasible region is diminished by the dual cuts, so that the CPDA would encounter a different problem upon reentry than it departed from—even though it re-entered at the same lattice point. This fact is undoubtedly the key to a formal convergence proof for the CPDA.

4. SUMMARY AND CONCLUSIONS

The CPDA is appealing for several reasons: (a) in primal algorithms, artificial variables must be used when the constraints (2) of (P) include "$\geqslant$" or "=" constraints, and the presence of such variables usually causes a significant increase in computational effort; (b) for highly degenerate problems, existing primal algorithms frequently encounter cycles which involve thousands of stationary iterations; (c) in all current dual algorithms, the objective function (1) of (P) must include only $c_j \geqslant 0$, or must be transformed into this form by finding upper bounds on the associated variables; (d) Glover's PPDA [6] (the only other primal-dual algorithm known to exist) can only solve minimization problems with some "$\geqslant$" and/or "=" constraints; and finally (e) the main drawback of all dual-based algorithms including Glover's PPDA is their inability to produce a "useful" solution before the integer optimal solution is reached. CPDA has this capability.

As regards computational efficiency, CPDA performed well in solving highly primal- and dual-degenerate fixed-charge problems. By incorporating an objective function cut, CPDA manages to delete many primal-dual-degenerate iterations, and eventually reaches the optimal solution in a small number of iterations. On the allocation problems, CPDA performed consistently, but did not dominate the other algorithms—as was the ca^e for the fixed-charge problems.

The general conclusion drawn from this research regarding the computational efficiency of the CPDA (in comparison with other IP algorithms) may be summarized as follows:

(a) Computation time is code dependent;

(b) Although an integer problem may be easily solved by one IP code, another code may prove a complete failure;

(c) CPDA can obtain a "useful" solution if the computation must be stopped prematurely. This is an important characteristic, since dual-based algorithms (and Glover's hybrid PPDA) do not have such a feature. No one code is uniformly better for all types of integer problems, although CPDA appears to perform extremely well on a set of difficult fixed-charge problems.

Two Caveats

First, we should note here that the standard test problems used in our computational testing—although difficult—are rather small. We are currently in the process of randomly generating larger problems for testing the algorithm, but preliminary results indicate that such problems—although larger—are relatively easy.

Second, we should acknowledge that branch-and-bound techniques for IP are generally considered superior to cutting-plane approaches. The problem with comparing these two techniques is that the criterion of interest in our study was the number of simplex iterations, rather than CPU time. Moreover, for ease of programming, the algorithm is written in SAS, which has remarkable matrix/vector handling capabilities, but is relatively inefficient in the sense of CPU time. We intend to further refine the algorithm (e.g., to incorporate advanced starts) and to reprogram it in PL-1. Once this is accomplished, we will be able to make comparisons regarding solution efficiency, by comparing our results with those obtained by IBM's MPSX, which is a branch-and-bound algorithm.

REFERENCES

[1] Austin, L.M., "The Bounded Descent Algorithm for Integer Programming," unpublished monograph, College of Business Administration, Texas Tech University, Lubbock, TX (1979).

[2] Glover, F., "A Pseudo Primal-Dual Integer Programming Algorithm," Journal of Research of the National Bureau of Standards, *71B*, 167-195 (1967).

[3] Gomory, R.E., "An Algorithm for Integer Solutions to Linear Programs," in *Recent Advances in Mathematical Programming*, R.L. Graves and P. Wolfe, Editors, 269-302 (McGraw-Hill, New York, 1963).

[4] Haldi, J., "25 Integer Programming Test Problems," Working Paper No. 43, Graduate School of Business, Stanford University, Stanford, CA (1964).

[4] Trauth, C.A. and R.E. Woolsey, "Integer Programming: A Study in Computational Efficiency," Management Science, *15*, 481-493 (1969).

[6] Wahi, P.N. and G.H. Bradley, "Integer Programming Test Problems," Report No. 28, Department of Administrative Science, Yale University, New Haven, CT (1969).

[7] Young, R.D., "A Simplified Primal (all-Integer) Integer Programming Algorithm," Operations Research, *16*, 750-782 (1968).

COMPUTING THE DISCOUNTED RETURN IN MARKOV AND SEMI-MARKOV CHAINS

Evan L. Porteus

Graduate School of Business
Stanford University
Stanford, California

ABSTRACT

This paper addresses the problem of computing the expected discounted return in finite Markov and semi-Markov chains. The objective is to reveal insights into two questions. First, which iterative methods hold the most promise? Second, when are iterative methods preferred to Gaussian elimination? A set of twenty-seven randomly generated problems is used to compare the performance of the methods considered. The observations that apply to the problems generated here are as follows: Gauss-Seidel is *not* preferred to Pre-Jacobi in general. However, if the matrix is reordered in a certain way and the author's row sum extrapolation is used, then Gauss-Seidel *is* preferred. Transforming a semi-Markov problem into a Markov one using a transformation that comes from Schweitzer does not yield improved performance. A method analogous to symmetric successive overrelaxation (SSOR) in numerical analysis yields improved performance, especially when the row-sum extrapolation is used only sparingly. This method is then compared to Gaussian elimination and is found to be superior for most of the problems generated.

1. INTRODUCTION

This paper addresses the problem of computing the unique N-vector v^* that satisfies $v^* = r + Pv^*$, where r is a given N-vector and P is nonnegative and has a spectral radius less than one. The primary intended application of this problem is finding the expected discounted return in infinite horizon, stationary, finite state, discrete time parameter, Markov and semi-Markov chains, although there are other interesting applications. For instance, this problem arises in solving Leontief input-output problems (Gale [8]) and when determining the value associated with a given basis when optimizing over Leontief substitution systems (see Veinott [34] and Koehler, Whinston, and Wright [14]). It also arises when using finite element and finite difference approximations to partial differential equations (see Varga [33], Young [37], Reid [26], and Fox [7]). In the latter case, P is usually symmetric and much of the modern advanced numerical analysis that deals with this problem exploits that fact, and is therefore not applicable to the Markov or semi-Markov chain case, since P is rarely symmetric then. Pertinent discussion of this problem in the context of Markov and semi-Markov chains can be found in references [2], [4-6], [12-14], [16], [19-25], [30-32], [35], and [36].

The objective of this paper is to shed light on two questions. First, which iterative methods, among a number of candidates, offer the most promise? Second, when are iterative methods preferred to Gaussian elimination? A set of twenty-seven Markov chain problems is generated and used to compare the various methods identified. There are nine problems with 50, 100, and 200 rows each. The rules used for generating the problems are described briefly in the appendix. The next section introduces some notation and sections 3-6 deal with the first question, regarding the promise of certain iterative methods. Section 7 deals with the second question and section 8 summarizes the conclusions of the paper.

2. NOTATION

We use the notation and conventions of Porteus [22], so only a partial exposition of them is given here. The L_∞ norm is used exclusively. The iterative methods we consider can be put into the form $\mathrm{v}^n = \tilde{r} + \tilde{P}\mathrm{v}^{n-1}$ when no extrapolations are used, where $\tilde{r}$ and $\tilde{P}$ may well be induced from an implicit transformation. The tildes are suppressed whenever clarity is preserved. We assume that $P \geqslant 0$ and $\rho(P) < 1$. Let α denote the vector of row sums of P, $\underline{\alpha}$ the minimum row sum, and $\bar{\alpha}$ the maximum row sum. In the Markov chain case, $\underline{\alpha} = \bar{\alpha}$, and, in the semi-Markov chain case, $\underline{\alpha} \leqslant \bar{\alpha}$. The sequence $\{\mathrm{v}^n\}$ *converges geometrically at the rate* β if $0 \leqslant \beta < 1$ and there exists a real number M such that $||\mathrm{v}^n - \mathrm{v}^*|| \leqslant \beta^n M$ for all n. The sequence *converges geometrically at the" rate"* β if it converges geometrically at the rate γ for all $\gamma > \beta$. The subradius of P is denoted by $\rho^*(P)$.

3. COMPARISON OF PRE-JACOBI AND GAUSS-SEIDEL ITERATION FOR MARKOV CHAIN PROBLEMS

It is well known that if Pre-Jacobi iteration starts with $\mathrm{v}^1 \geqslant \mathrm{v}^0$ (so that $\mathrm{v}^{n+1} \geqslant \mathrm{v}^n \geqslant \ldots \geqslant \mathrm{v}^0$) then Gauss-Seidel iteration yields $\tilde{\mathrm{v}}^n \geqslant \mathrm{v}^n$ for all n, where $\{\tilde{\mathrm{v}}^n\}$ denotes the sequence generated by Gauss-Seidel. A common interpretation of this result is that Gauss-Seidel is faster than Pre-Jacobi. We shall argue that this interpretation is misleading at best. Starting with a specified $\epsilon > 0$, the object of an iterative method here is to obtain an estimate $\mathrm{v}_{\mathrm{est}}$ of v^* such that $||\mathrm{v}_{\mathrm{est}} - \mathrm{v}^*||/||\mathrm{v}^*|| \leqslant \epsilon$ is guaranteed. (We used $\epsilon = 10^{-6}$ in all of the numerical work discussed in this paper.) A stopping approach consists of a technique for constructing $\mathrm{v}_{\mathrm{est}}$ from the data generated by the method and a rule for stopping the computations that guarantees $||\mathrm{v}_{\mathrm{est}} - \mathrm{v}^*||/||\mathrm{v}^*|| \leqslant \epsilon$.

There are two important alternative approaches that we wil' discuss. The first uses v^n as the estimate of v^* and stops when $||\mathrm{v}^n - \mathrm{v}^{n-1}||/||\mathrm{v}^n|| \leqslant \epsilon(1-\bar{\alpha})/[\bar{\alpha}(1+\epsilon)]$. This rule derives from the classical bound

$$||\mathrm{v}^n - \mathrm{v}^*|| \leqslant ||P|| \cdot ||\mathrm{v}^n - \mathrm{v}^{n-1}||/(1 - ||P||),$$

since $||P|| = \bar{\alpha}$.

The second approach selects

$$\mathrm{v}_{\mathrm{est}} = \mathrm{v}^n + \bar{\alpha}(a_n + b_n)/[2(1-\bar{\alpha})],$$

where $a_n := \min_i (\mathrm{v}_i^n - \mathrm{v}_i^{n-1})$ and $b_n := \max_i (\mathrm{v}_i^n - \mathrm{v}_i^{n-1})$ and stops when $b_n - a_n \leqslant 2\epsilon(1-\bar{\alpha})$ $||\mathrm{v}_{\mathrm{est}}||/[\bar{\alpha}(1+\epsilon)]$. This rule derives from the bounds of Porteus [24]. It improves on the first because $b_n - a_n \leqslant 2||\mathrm{v}^n - \mathrm{v}^{n-1}||$ always holds, often with a substantial difference since $\mathrm{v}^n \geqslant \mathrm{v}^{n-1}$ means that $a_n \geqslant 0$.

At this point, one might well ask what the above has to do with a comparison between Pre-Jacobi and Gauss-Seidel. If the first stopping approach is used, then clearly

$||\tilde{v}^n - v^*|| \leqslant ||v^n - v^*||$, giving favor to Gauss-Seidel. However, if the second approach is used, then favor can swing to Pre-Jacobi. For Markov chain problems, Pre-Jacobi is equivalent to relative value iteration, which converges geometrically at the "rate" $\rho^*(P)$. (The equivalence holds only when the second stopping approach is used.) Gauss-Seidel converges geometrically at the "rate" $\rho(\tilde{P})$, where $\tilde{P}$ denotes the implied transition matrix. Thus, if $\rho^*(P) < \rho(\tilde{P})$, which is very possible, Pre-Jacobi has a better guaranteed speed of convergence than does Gauss-Seidel.

To assess the extent to which the observations above are borne out in examples, we applied both Pre-Jacobi and Gauss-Seidel with each of the stopping approaches mentioned to the twenty-seven problems generated. For Gauss-Seidel, the second stopping approach used the appropriate specialization of the results in Porteus [24]: $\tilde{v}^n \geqslant \tilde{v}^{n-1}$ implies that the selection is

$$v_{est} = \tilde{v}^n + .5\,[\underline{\alpha}\,a_n/(1-\underline{\alpha}) + \bar{\alpha}\,b_n/(1-\bar{\alpha})],$$

which guarantees that $||v_{est} - v^*|| \leqslant .5\,[\bar{\alpha}\,b_n/(1-\bar{\alpha}) - \underline{\alpha}\,a_n/(1-\underline{\alpha})]$. The trivial scalar extrapolation was used after each iteration of Pre-Jacobi to reduce roundoff errors. Doing so has no effect on the results in the absence of roundoff errors. However, we could not use that extrapolation when applying Gauss-Seidel because it affects the results. Indeed, it can yield a divergent method. Provided the minimum and maximum row sums of the Gauss-Seidel matrix are not too close (which was the case in all of our examples), the effect of roundoff errors is significantly smaller than the effect when Pre-Jacobi is applied. Thus, the comparison is not significantly affected by the fact that measures were taken to control the effect of roundoff errors for Pre-Jacobi but none were for Gauss-Seidel.

The results are summarized in Table 1. Some of the detailed results appear later in Table 4, but comprehensive details are omitted here because there was so little variation in the results within each method. The results show that when the first stopping approach is used, Gauss-Seidel is preferred to Pre-Jacobi, but that the opposite is the case with the second approach. Indeed, under the second approach, Pre-Jacobi always required fewer iterations with the differences ranging from 6, 14, and 17 up to 48. Incidentally, using Gauss-Seidel, the second stopping approach always required 4 or 5 fewer iterations than the first.

TABLE 1 — *Number of Iterations Required by Pre-Jacobi Iteration and Gauss-Seidel Iteration Using Two Alternative Stopping Approaches, for Twenty-Seven Problems*

Stopping Approach	Iterative Method	Number of Iterations		
		Minimum	Median	Maximum
First	PJ	112	123	128
	GS	60	66	83
Second	PJ	15	32	72
	GS	56	62	78

To see that the differences can become more pronounced when the row-sums are close to one, the probabilities in the nine problems with 50 rows each were scaled up so that their row-sums changed from .9 to .99, and the four combinations were applied. Pre-Jacobi with the second stopping approach required approximately twice as many iterations whereas, not surprisingly, the other three combinations required more than ten times as many. The second stopping approach will be used exclusively, henceforth.

4. THE EFFECT OF REORDERING AND EXTRAPOLATIONS

The previous section gives examples in which Pre-Jacobi is preferred to Gauss-Seidel. However, no reorderings or extrapolations were carried out in these examples (except that which was done to control roundoff error). In Porteus and Totten [25] and Porteus [22], evidence was given that both reordering and extrapolating can have a beneficial effect when Gauss-Seidel is used. Therefore, both of these modifications were applied to the twenty-seven test problems. The minimum remaining row-sum (reordering) method and the row-sum extrapolation were used. Both were introduced in Porteus [22], where they outperformed their identified competitors on the numerical examples considered. The numerical results here are summarized in Table 2. To attempt to make the comparisons between methods fair, the results are reported in terms of normalized (Pre-Jacobi) iterations, which account for the additional multiplications and divisions needed (such as for the extrapolations). Some of the detailed results appear in Table 4.

TABLE 2 — *Median Number of Normalized Iterations Required by Gauss-Seidel with and without Reordering and/or Extrapolating*

Iterative Method	Reordering?	Extrapolating?	Number of Rows			
			50	100	200	Overall
PJ	No	No*	40	30	32	32
GS	No	No	63	62	60	62
GS	Yes	No	43	36	35	36
GS	No	Yes	30.7	29.6	29.3	30.6
GS	Yes	Yes	17.8	18.4	20.3	18.5

*Except for control of roundoff error, as previously indicated.

The results show that both reordering and extrapolating provide beneficial effects on Gauss-Seidel for the examples considered. Indeed, when both modifications were used, Gauss-Seidel outperformed Pre-Jacobi in every instance, reversing the order of preference observed in the previous section. This preference, for Gauss-Seidel (with reordering and extrapolating) over Pre-Jacobi, appears to be valid even when the row-sums are very close to one, as it was observed in the numerical examples in Porteus [22] in which row-sums were .9, .99, .999, and .9999.

5. THE IMPLIED EQUAL ROW SUM TRANSFORMATION

This transformation converts a semi-Markov chain problem (with unequal row-sums) into an equivalent Markov chain problem (with equal row-sums). It was introduced by Schweitzer [27], discussed by Porteus [21] and by van Nunen and Stidham [32] with a computational context in mind, applied by Lippman [15] to queuing optimization models, and studied by Serfozo [29] with some general Markov decision processes in mind.

It is called the implied equal row-sum transformation here because, computationally, it is desirable to implement it in the following manner. Assume without loss of generality that $p_{ii} = 0$ for all i. Let $w_i := (1 - \bar{\alpha})/(1 - \alpha_i)$, where α_i and $\bar{\alpha}$ are as defined in Section 2. The resulting iterative method (corresponding to applying Pre-Jacobi in conjunction with the transformation) is as follows:

$$\tilde{v}^n = r + Pv^{n-1}$$

$$v_i^n = w_i \tilde{v}_i^n + (1 - w_i)\, v_i^{n-1} \textit{ for all } i.$$

The implied row sums are all equal to $\bar{\alpha}$.

If one starts with an unequal row-sum problem, then Pre-Jacobi will converge at the "rate" of the radius (of P) whereas if the implied equal row-sum transformation is applied, yielding an induced transition matrix $\tilde{P}$, then Pre-Jacobi will converge at the "rate" of the subradius of $\tilde{P}$. Thus, it is not necessarily clear that this method will converge faster than Pre-Jacobi applied to the original problem.

To observe the potential of this method, it was applied to two sets of twenty-seven semi-Markov chain problems. Both sets were derived from the original twenty-seven Markov chain problems. Each time Gauss-Seidel is applied to a Markov chain problem using a particular ordering of the states, there is an induced transition matrix with unequal row-sums. Such a transition matrix can be interpreted as representing a semi-Markov chain problem. The method discussed in this section is then applied to that problem. The first set of semi-Markov chain problems derived from applying Gauss-Seidel to the original matrices and the second set from applying Gauss-Seidel to the reordered matrices. In the first set, the implied equal row-sum (IERS) method reduced the median number of normalized iterations from 62 to 47.1. However, the figures are 36 and 64.9, respectively, for the second set. A partial explanation of this reversal in performance may involve $\bar{\alpha} - \underline{\alpha}$, the difference between the maximum and minimum row sums. When this difference is large, IERS entails large implied diagonal entries, which correspond to a large subradius. In the first set of problems, this difference was relatively small, with a median value of .19, whereas it was larger in the second set of problems, where the median was .31. Thus, it seems doubtful that IERS will yield improved results in general. Indeed, when the row sum extrapolation was used instead of IERS, the median number of normalized iterations was reduced from 47.1 to 30.6 for the first set of problems and from 64.9 to 18.5 for the second. However, IERS may yield improved results in problems in which $\bar{\alpha} - \underline{\alpha}$ is quite small. Furthermore, in semi-Markov *decision* problems, it may well be valuable in selecting policies during the early stages of an algorithm, since it equalizes transition times and reduces the sensitivity of the policy improvement step to the estimate of the optimal return function.

6. REVERSIBLE GAUSS-SEIDEL

Reversible Gauss-Seidel (RGS) is simply an analogous method to what is called SSOR (symmetric successive overrelaxation) in the numerical analysis literature (see Varga [33] or Young [37], for example). It consists of using Gauss-Seidel in a specified order during the first pass of an iteration and in the reverse order during the second pass. An important advantage of this method, as presented by Conrad and Wallach [1], is that with modest additional storage, a substantial number of operations can be saved. For instance, if the original matrix has an equal number of nonzero entries in its lower and upper triangular parts, then the second pass requires only about one-half of a normalized iteration. Furthermore, if no extrapolation is made after an iteration, the first pass of the subsequent iteration also requires only about one-half of a normalized iteration. (If an extrapolation is made, then the first pass of the next iteration requires a full normalized iteration.)

When the row sum extrapolation was used, RGS required a median of 24.3 normalized iterations over the standard twenty-seven problems and 16.7 when the matrix was reordered. Note that each iteration requires roughly 1½ normalized iterations, because an extrapolation is made after each iteration. If an extrapolation is likely to have little effect on the current iterate, it might be useful to skip the extrapolation, so that the next iteration would require merely one normalized iteration. A modification based on this idea was tried. Let v^n denote the nth iterate and $\tilde{v}^{n+1}$ the result after applying the two passes of the iteration, before any extrapolation. The rule used was: if

$$\min_i (\tilde{v}_i^{n+1} - v_i^n) < 0 < \max_i (\tilde{v}_i^{n+1} - v_i^n),$$

then do not extrapolate. This modification brought the median number of normalized iterations down to 19.3, and to 15.5 for the reordered matrices, which represented the best performance obtained over all methods tested during this study. Detailed results appear in Table 4.

7. COMPARISON OF ITERATIVE METHODS AND GAUSSIAN ELIMINATION

It is well known that iterative (indirect) methods, such as Pre-Jacobi iteration, take less computational effort to obtain a practical solution to our problem than do direct methods, such as Gaussian elimination, as long as the number of rows is sufficiently large, *provided* that P is fully dense ($p_{ij} > 0$ for all i and j). It is plausible to speculate that the result is true when P is sparse as well. Our numerical results will support this claim. Indeed, they will give insight into the approximate combination of problem size and density for which an iterative method is preferred to Gaussian elimination.

The effectiveness of Gaussian elimination is significantly affected by the rule used for selecting the pivot elements. A rule which "has proved to be satisfactory over a very wide range of problems," according to Reid [26], is the one proposed by Markowitz [17]. It iteratively selects as the next pivot element one that satisfies a numerical stability criterion and minimizes the product of nonzeroes in its column (over other unanalyzed rows) and nonzeroes in its row (over other unanalyzed columns). In general, all nonzeroes in the *unanalyzed submatrix* (the matrix composed of unanalyzed rows and columns) are considered as possible pivot elements. In each of the problems generated here, the original matrix I-P is a diagonally dominant M-matrix (see Varga [33] and Young [37] for convenient discussions of M-matrices in the context of iterative methods). A result by Fan [3] and an observation by Meijerink and van der Vorst [18] guarantee that after pivoting on any diagonal element of a diagonally dominant M-matrix, the remaining unanalyzed submatrix will also be a diagonally dominant M-matrix. Thus, only diagonal entries were allowed to be pivot elements in our numerical work. This rule guarantees that numerical stability will be maintained regardless of which diagonal element is selected as the pivot element and eliminates the need to compute the indicated product of nonzeroes for potential pivot elements except for the diagonal elements.

The object of the numerical work in this section is to compare Gaussian elimination with the best found iterative method, which was reversible Gauss-Seidel with the matrix reordered and using the row sum extrapolation occasionally, as specified by the rule discussed in the previous section. For convenience, we refer to this method simply as RGS in this section.

With the accuracy criterion used to stop RGS (see Section 3) and the computer used, both RGS and Gaussian elimination yielded results of comparable accuracy. To be conservative on the side of favoring Gaussian elimination, the multiplications needed to select the pivot elements were not counted. Furthermore, we assumed that a data structure was used with Gaussian elimination that required multiplications to be carried out only when two nonzeroes were to be multiplied together. This assumption also favors Gaussian elimination.

The results appear in Table 3. The first two columns indicate parameter values used to generate the problems: the maximum probability size and the maximum number of nonzero probabilities in each row. The third column refers to the average number of nonzero probabilities in each row (averaged over the rows for a given problem). The range shown corresponds to the minimum and maximum such averages over the three problems represented in that row. Before discussing the results, it is worth mentioning that the Markowitz ordering gave substantially better results for Gaussian elimination than was obtained with either the original, essentially random, ordering or the ordering used by RGS. Indeed, use of the original ordering required anywhere from 1.6 to 7.3 times as much work, with an average of 4.5, as the Markowitz ordering.

TABLE 3 — *Number of Normalized Iterations Required by RGS and Gaussian Elimination (GE)*

Number of Nonzero Probabilities per Row			Number of Rows					
Maximum Probability	Maximum	Actual Range	50		100		200	
			RGS	GE	RGS	GE	RGS	GE
.9	3	2.4-2.5	19.4	5.5	15.6	9.6	18.1	29.3
.75	3	2.7	16.6	6.4	13.8	13.1	17.1	41.9
.9	5	2.6-2.7	18.9	8.9	15.6	12.8	23.1	43.5
.9	10	2.7-2.8	16.6	10.0	14.2	16.2	17.5	37.4
.6	3	2.9	15.7	6.7	15.7	17.6	15.6	54.4
.63	5	3.4-3.5	15.5	11.1	12.1	29.5	14.6	78.1
.54	10	3.9-4.1	13.1	16.1	13.4	35.9	13.5	101.8
.36	5	4.7-4.8	11.5	19.7	12.8	52.2	12.6	151.4
.18	10	9.5-9.6	11.0	31.3	10.9	81.8	11.0	268

The results support the claim that for problems of a given density (number of nonzero elements per row) an iterative method (RGS in this case) will outperform GE for problems with sufficiently many rows. Indeed, they support the hypothesis that, for fixed density, the number of normalized iterations (total number of operations, respectively) required by RGS is a constant (linear, respectively) function of the number of rows. Similarly, for Gaussian elimination, the supported relationships are quadratic and cubic functions for normalized iterations and total work, respectively.

The numerical results also support the hypothesis that for problems of a given size (number of rows), an iterative method will outperform GE for problems with sufficiently high density. For Gaussian elimination, higher density is apt to imply a larger number of fill-ins (newly created nonzeroes) of the unanalyzed submatrices as the partial pivoting is carried out. Such fill-ins tend to compound themselves and cause the number of multiplications required to

TABLE 4 — *Number of Normalized Iterations Required by Various Algorithms on Each Problem*

Problem			Algorithm											
NR	NNZ	MAXP	GE	PJ	GS11	GS21	GS12	GS22	RG12	RG13	RG22	RG23	GI11	GI21
50	3	.9	5.5	72	78	46	37.2	28.9	28.7	20.6	24.8	19.4	66.9	62.7
100	3	9	9.6	36	62	34	39.3	23.7	29.4	23.7	17.7	15.6	51.0	76.5
200	3	.9	29.3	42	62	29	35.9	23.4	34.4	24.5	23.6	18.1	58.7	88.0
50	3	.75	6.4	52	74	41	34.0	23.0	26.8	20.6	19.4	16.6	57.8	61.9
100	3	.75	13.1	33	61	32	35.3	18.9	24.7	22.3	16.4	13.8	48.1	78.3
200	3	.75	41.9	36	62	35	31.1	22.9	26.5	21.3	21.9	17.1	49.3	71.3
50	5	.9	8.9	42	56	33	42.5	23.2	28.3	21.4	22.6	18.9	66.5	85.8
100	5	.9	12.8	49	67	32	39.8	24.5	31.7	21.4	18.3	15.6	60.9	88.6
200	5	9	43 5	40	57	26	40.8	27.1	37.3	26.1	32.1	23.1	57.6	90.5
50	10	.9	10.0	46	65	38	31.2	16.1	28.5	19.6	17.0	16.6	56.3	75.5
100	10	.9	16.2	41	63	31	29.6	17.4	36.1	26.4	17.0	14.2	50.5	76.4
200	10	.9	37 4	36	59	26	35.4	20.3	26.6	21.5	18.1	17.5	49.5	90.8
50	3	.6	6.7	40	70	43	30.6	21.2	24.5	19.3	19.5	15.7	48.5	52.5
100	3	.6	17.6	30	61	36	32.0	18.5	22.5	19.2	16.7	15.7	45.8	71.5
200	3	6	54.4	32	62	36	29.3	18.5	24.3	20.2	18.3	15.6	47.1	68.7
50	5	63	11.1	30	60	37	30.7	17.8	21.6	17.7	16.4	15.5	42.6	62.0
100	5	.63	29.5	29	63	36	25.7	16.6	19.6	16.8	14.5	12.1	44.1	64.9
200	5	63	78.1	28	57	36	24.2	20.3	21.5	17.8	16.5	14.6	43.8	63.1
50	10	.54	16.1	25	63	44	22.2	14.7	17.9	16.6	12.7	13.1	37.5	54.9
100	10	.54	35.9	25	62	38	25.9	18.4	17.8	16.5	13.0	13.4	36.1	62.3
200	10	54	101.8	25	60	35	26.1	17.3	19.6	17.6	14.3	13.5	39.0	71.6
50	5	36	19.7	23	63	45	18.0	15.5	15.7	14.5	12.7	11.5	33.9	46.0
100	5	.36	52.2	22	62	43	18.0	16.8	17.1	14.4	13.0	12.8	36.4	47.3
200	5	.36	151.4	22	58	38	20.3	17.9	17.4	15.5	12.7	12.6	38.7	53.2
50	10	.18	31.3	15	63	51	14.3	10.9	13.2	12.0	11.0	11.0	27.6	34.3
100	10	.18	81.8	15	62	47	16.5	12.0	13.2	13.2	10.9	10.9	29.8	40.9
200	10	.18	268.0	16	61	47	15.4	14.3	13.3	14.3	11.0	11.0	27.6	39.8
Median			29.6	32	62	36	30.6	18.5	24.3	19.3	16.7	15.5	47.1	64.9
Mean			44.0	33.4	62.7	37.6	28.9	19.3	23.6	19.1	17.1	15.0	46.4	65.9

Symbols used in table:

NR	Number of rows
NNZ	Maximum number of nonzero elements per row
MAXP	Maximum probability in each row
GE	Gaussian elimination, using Markowitz ordering
PJ	Pre-Jacobi using trivial scalar extrapolation for roundoff error control
GS11	Gauss-Seidel, using original ordering and no extrapolations
GS21	Gauss-Seidel, using reordered matrix and no extrapolations
GS12	Gauss-Seidel, using original ordering and row sum extrapolation
GS22	Gauss-Seidel, using reordered matrix and row sum extrapolation
RG12	Reversible Gauss-Seidel, using original ordering and row sum extrapolation
RG13	Reversible Gauss-Seidel, using original ordering and modified row sum extrapolation
RG22	Reversible Gauss-Seidel, using reordered matrix and row sum extrapolation
RG23	Reversible Gauss-Seidel, using reordered matrix and modified row sum extrapolation
GI11	IERS after Gauss-Seidel on original ordering, using trivial scalar extrapolation for roundoff error control
GI21	IERS after Gauss-Seidel on reordered matrix, using trivial scalar extrapolation for roundoff error control

increase significantly. On the other hand, if anything, RGS appears to require fewer normalized iterations as the density increases. An explanation of this observation may have something to do with the following. RGS appears to converge at least as fast as Pre-Jacobi which (when applied to Markov chain problems) converges at the "rate" of the subradius of the transition matrix. Morton and Wecker [20] discuss some upper bounds on this subradius. These are called n-Hajnal measures because they are based on the work of Hajnal [9, 10]. These upper bounds are likely to decrease as the density is increased, for problems of the type generated here.

8. CONCLUSION

This paper has tried to provide some insights into the questions of which iterative methods hold the most promise for computirg the expected discounted return in finite Markov and semi-Markov chains and when iterative methods are preferred to Gaussian elimination. For the first question, the results are inconclusive for two reasons. One, not all, competing methods were tested. For example, Verkhovsky's [35] method was not tested. Two, specific rules were used to randomly generate the twenty-seven problems used to test the methods. It is unclear that the results found here will apply to other problems. Nevertheless, there were some clear observations that applied to the problems that were generated. Gauss-Seidel is not preferred to Pre-Jacobi for Markov chain problems in general, although it is preferred if the matrix is reordered and the row-sum extrapolation is used. The implied equal row-sum transformation did not yield improved results. Reversible Gauss-Seidel did yield slightly improved performance, provided that the matrix was reordered and the row-sum extrapolation was used. A further slight improvement was obtained when a rule for only applying the extrapolations occasionally was used. Whether these improvements are worth the extra coding needed to implement them remains to be seen. However, it appears that iterative methods are preferred to Gaussian elimination for some problems with as few as 50 rows and for many problems that have 200 rows or more.

APPENDIX: PROBLEM GENERATION

All twenty-seven problems generated were Markov chain problems with a discount factor of .9 ($\underline{\alpha} = \bar{\alpha} = .9$). Three other parameters were used to generate the problems. Their values for each problem are shown in Table 4. NR is simply the number of rows. NNZ and MAXP were used to generate the probabilities in a row.

The locations of the nonzero probabilities in each row were generated randomly. Working sequentially, a location was chosen from a discrete uniform distribution over the positions not yet having a nonzero element. The values of the elements were generated from a discrete uniform distribution on the interval $[.001, q]$ where $q := \min$ (MAXP, REM), and REM is the discount factor less the cumulative sum of probabilities already generated in that row. If the resulting cumulative row sum did not equal the discount factor and the maximum number of nonzero elements (NNZ) had not yet been attained for that row, then another location was generated. Because a minimum element size (equal to .001) was used, the maximum number of nonzero elements was not always attained. The last nonzero element generated in each row was adjusted so that the row sum would equal the discount factor.

The immediate returns (making up the vector r) were generated as follows. The first ten were set equal to $10n$, where n denotes the row number. The last was set equal to zero, and the remaining ones were generated from a continuous uniform distribution on the interval (0, 1). The rationale behind this rule is outlined in Porteus [22].

REFERENCES

[1] Conrad, V. and Y. Wallach, "A Faster SSOR Algorithm," Numerische Mathematik, *27*, 371-372 (1977).

[2] Denardo, E., "Contraction Mappings in the Theory Underlying Dynamic Programming," SIAM Review, *9*, 165-177 (1967).

[3] Fan, K., "Note on M-Matrices," Quarterly Journal of Mathematics, *11*, 43-49 (1960).

[4] Federgruen, A., P. Schweitzer and H. Tijms, "Contraction Mappings Underlying Undiscounted Markov Decision Problems," Journal of Mathematical Analysis and Applications, *65*, 711-730 (1978).

[5] Federgruen, A. and P. Schweitzer, "A Survey of Asymptotic Value-Iteration for Undiscounted Markovian Decision Processes," in *Recent Developments in Markov Decision Processes*, R. Hartley, L. Thomas and D. White, Editors (Academic Press, New York, 1980).

[6] Fox, B., "Reducing the Number of Multiplications in Iterative Processes," Acta Informatica, *3*, 43-45 (1974).

[7] Fox, L., "Finite-Difference Methods for Elliptic Boundary-Value Problems," in *The State of the Art in Numerical Analysis*, D. Jacobs, Editor (Academic Press, London, New York, 1977).

[8] Gale, D., *The Theory of Linear Economic Models* (McGraw-Hill, New York, 1960).

[9] Hajnal, J. "The Ergodic Properties of Nonhomogeneous Markov Chains," Proceedings of the Cambridge Philosophical Society, *52*, 67-77 (1956).

[10] Hajnal, J., "Weak Ergodicity in Nonhomogeneous Markov Chains," Proceedings of the Cambridge Philosophical Society, *54*, 233-246 (1958).

[11] Hitchcock, D. and J. MacQueen, "On Computing the Expected Discounted Return in a Markov Chain," Naval Research Logistics Quarterly, *17*, 237-241 (1970).

[12] Howard, R., *Dynamic Programming and Markov Processes*, (John Wiley, New York, 1960).

[13] Jewell, W., "Markov-Renewal Programming. I: Formulation, Finite Return Models," Operations Research, *11*, 938-948 (1963).

[14] Koehler, G., A. Whinston and G. Wright, *Optimization over Leontief Substitution Systems* (North-Holland, Amsterdam, 1975).

[15] Lippman, S., "Applying a New Device in the Optimization of Exponential Queueing Systems," Operations Research, *23*, 687-710 (1975).

[16] MacQueen, J., "A Modified Dynamic Programming Method for Markovian Decision Problems," Journal of Mathematical Analysis and Applications, *14*, 38-43 (1966).

[17] Markowitz, H., "The Elimination Form of the Inverse and Its Application to Linear Programming," Management Science, *3*, 255-269 (1957).

[18] Meijerink, J. and H. Van der Vorst, "An Iterative Solution Method for Linear Systems of which the Coefficient Matrix Is a Symmetric M-Matrix," Mathematics of Computatior. *137*, 148-162 (1977).

[19] Morton, T., "On the Asymptotic Convergence Rate of Cost Differences for Markovian Decision Processes," Operations Research, *19*, 244-248 (1971).

[20] Morton, T. and W. Wecker, "Discounting, Ergodicity, and Convergence for Markov Decision Processes," Management Science, *23*, 890-900 (1977).

[21] Porteus, E., "Bounds and Transformations for Finite Markov Decision Chains," Operations Research, *23*, 761-784 (1975).

[22] Porteus, E., "Improved Iterative Computation of the Expected Discounted Return in Markov and Semi-Markov Chains," Zeitschrift fuer Operations Research, *24*, 155-170 (1980).

[23] Porteus, E., "Overview of Iterative Methods for Discounted Finite Markov and Semi-Markov Decision Chains," in *Recent Developments in Markov Decision Processes*, R. Hartley, L. Thomas and D. White, Editors (Academic Press, New York, 1980).

[24] Porteus, E., "Some Bounds for Discounted Sequential Decision Processes," Management Science, *18*, 7-11 (1971).

[25] Porteus, E. and J. Totten, "Accelerated Computation of the Expected Discounted Return in Markov Chain," Operations Research, *26*, 350-358 (1978).

[26] Reid, J., "Sparse Matrices," in *The State of the Art in Numerical Analysis*, D. Jacobs, Editor (Academic Press, New York, 1977).

[27] Schweitzer, P., "Iterative Solution of the Functional Equations of Undiscounted Markov Renewal Programming," Journal of Mathematical Analysis and Applications, *34*, 495-501 (1971).

[28] Schweitzer, P. and A. Federgruen, "Geometric Convergence of Value-Iteration in Multichain Markov Renewal Programming," Advances in Applied Probability, *11*, 188-217 (1979).

[29] Serfozo, R., "An Equivalence between Continuous and Discrete Time Markov Decision Processes," Operations Research, *27*, 616-620 (1979).

[30] Van Hee, K., A. Hordijk and J. Van der Wal, "Successive Approximations for Convergent Dynamic Programming," in *Markov Decision Theory*, H. Tijms and J. Wessels, Editors (Mathematical Centre Tract 93, Amsterdam, 1977).

[31] Van Nunen, J., "A Set of Successive Approximation Methods for Discounted Markovian Decision Problems," Zeitschrift fuer Operations Research, *20*, 203-208 (1976).

[32] Van Nunen, J. and S. Stidham, Jr., "Action-Dependent Stopping Times and Markov Decision Processes with Unbounded Rewards," OR Report No. 140, North Carolina State University, Raleigh, NC (1978).

[33] Varga, R., Matrix Iterative Analysis (Prentice-Hall, Englewood Cliffs, New Jersey, 1962).

[34] Veinott, A., Jr., "Extreme Points of Leontief Substitution Systems," Linear Algebra and Its Applications, *1*, 181-194 (1968).

[35] Verkhovsky, B., "Smoothing System Design and Parametric Markovian Programming," in *Markov Decision Theory*, H. Tijms and J. Wessels, Editors (Mathematical Centre Tract 93, Amsterdam, 1977).

[36] White, D., "Dynamic Programming, Markov Chains and the Method of Successive Approximations," Journal of Mathematical Analysis and Applications, *6*, 373-376 (1963).

[37] Young, D., *Iterative Solution of Largc Linear Systems* (Academic Press, New York, 1971).

EXACT ANALYSIS OF A TWO-ECHELON INVENTORY SYSTEM FOR RECOVERABLE ITEMS UNDER BATCH INSPECTION POLICY

Kripa Shanker

Industrial and Management Engineering Programme
Indian Institute of Technology
Kanpur, India

ABSTRACT

We investigate a two-echelon (base-depot) inventory system of recoverable (repairable) items. The arrivals of demand at the bases are in a Poisson manner and the order sizes are random. The failed units can be repaired either at the base or at the depot, and the units beyond economic repair are condemned. Inspection of the failed units is carried out in the batches they arrive, that is, arrival batches are not broken up. The exact expressions for stationary distribution of depot inventory position, and of the number of backorders, on-hand inventory, in-repair inventory at all locations are derived under the assumptions of constant repair and lead times. Special cases of complete recoverability, nonrecoverability, and of the unit order size are also discussed.

1. SYSTEM

Consider a two-echelon inventory system consisting of a set of bases (lower echelon) and a central depot (upper echelon). Each location, in addition to being an item stocking point, has facilities to perform repairs. Figure 1 shows the schematic diagram of the system under consideration. The system demands are generated at bases. Customers while placing requisitions for certain number of items, turn in a like number of failed units. The item stocked in this system is recoverable. Upon failures, units are returned to base where a decision is made e'ther to remove (condemn) the units from the system or to perform repairs on them either at the base itself or at the depot in order to restore the units to a serviceable condition. The decision to repair or to condemn the failed units is based on the degree and the nature of failure, the repair facilities available, and the economics involved. Once an item is designated as recoverable, it is presumably more economical to repair the item than it is to dispose of it and replace it with a new item.

The system is further characterized by the following descriptors:

(1) Arrival Pattern and Order Size. The failures which generate the system demands occur in a Poisson manner with known and constant rate. Upon such failures, a random number of units is demanded for replacement.

(2) Resupply at Bases and Depot. The bases are resupplied as necessary only by the depot. That is, lateral supply among the bases is not allowed. The procurement of items from

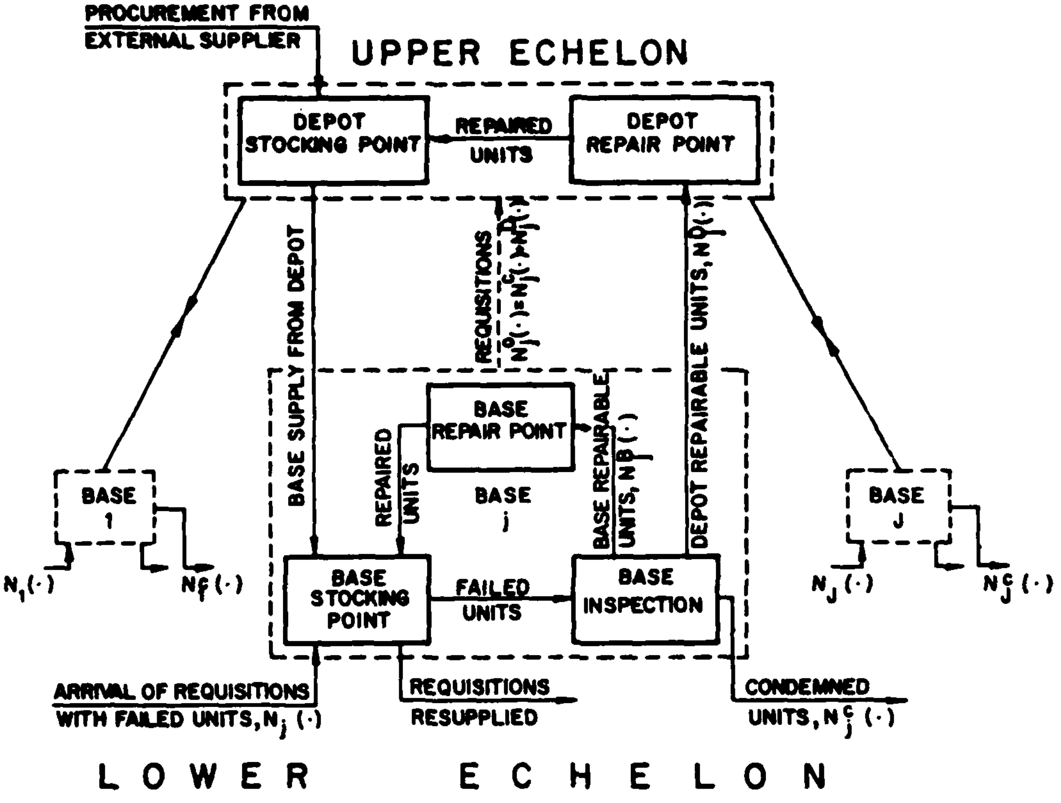


FIGURE 1. Two-echelon inventory system for recoverable items

the external supplier to make up for the condemnations i~~s~~ made only by the depot. Each location, in addition, receives supplies from its repair station.

(3) Inspection Policy at Bases. The inspection of failed units is carried out at the same base where the requisition for replacement arrived immediately after failure. Two inspection policies suggest themselves. The first policy, Batch Inspection Policy, determines if a batch of failed units as a whole is either base repairable, depot repairable or condemnable. From a practical view point, this policy represents situations where the units of a batch (or a module) fail simultaneously for the same reason and the extent of damage is the same for all units within the batch. In the second policy, Unit Inspection Policy, each failed unit in a batch is inspected independently to determine whether the unit is base repairable, depot repairable or condemnable. This policy represents the situations where units failed under difficult conditions but are submitted in a batch for replacement, or the situations where decision is very critical and calls for inspection of each individual unit in a batch. In this paper we shall consider Batch Inspection Policy. It is further assumed that the inspection takes negligible time.

(4) Procurement Policies. The bases use an $(s - 1, s)$ policy for procurement of units from the depot. The depot procurement policy is a general continuous review (s, S) policy. Policies at all locations are in terms of inventory position defined to be the sum of on-hand, on-order and in-repair inventories minus any backorders.

(5) Backlogging. Demands occurring when a location is out of stock are backlogged. Further, partial backlogging of demands is allowed. For example, upon arrival of a requisition if the base does not have the number of units demanded, then all the units on hand are supplied while the balance is backlogged. The respective base, however, accepts the whole batch of failed units and starts the inspection. Similarly, partial backlogging is done at the depot for the base demands.

(6) Repairs. There are ample repair facilities at each location. The repair times at the bases and at the depot are deterministic and independent of arrival process and of the number of units in repair. Furthermore, the depot repair time is the same for all the units received from all the bases. Repairs at all locations are carried out in a continuous manner—that is, no batching is done. Upon completion of repairs, the units immediately join the stock of serviceable items.

(7) Shipping and Lead Times. The time to ship a depot-repairable unit to the depot from a base is assumed to be negligible. In reality, it can be absorbed in the depot repair time. The order-and-ship time of unit from depot to a base is assumed to be deterministic. Similarly, order-and-ship time of an order from external supplier to the depot is also deterministic.

The objective function of the total expected cost and several measures of performance of the inventory system are related to the stationary distribution of depot inventory position, and number of backorders, on-hand inventory, and in-repair inventory at different locations. Our objective in this paper is to determine the exact expressions for stationary distribution of these inventory levels.

A fundamental work on two-echelon inventory system was the development of METRIC by Sherbrooke [7] for a completely conservative system that does not allow item condemnations. He considered the problem of allocating several units among a depot and several bases in order to minimize the total expected number of backorders at bases within the limitations of a budget. The resulting expressions for stationary distribution of backorder, however, are approximate. A variation of the METRIC model was introduced by Simon [8] to obtain the exact expressions for stationary distribution of backorders, allowing condemnations. Simon's analysis was limited to the case of unit order size. For unit order size, Muckstadt [2, 3] has presented analyses of the system where bases use (s, S) and (r, Q) procurement policies. Shanker [5] has analyzed the case of random order size with condemnations; the condemnation rates, however, were assumed to be the same at all bases. Our present analysis is more general than previous works in that random order size and condemnations both are considered. It, however, is less general than METRIC in that repair times here are assumed deterministic.

2. MODEL

We consider a two-echelon system as depicted in Figure 1 with J bases, the bases numbered from 1 to J and the depot indexed as 0. The failures which generate the system demands occur in a Poisson manner with known rate λ_j at base $j (j = 1, 2, \ldots, J)$. Upon such failures at base j, the number of units demanded, or equivalently, the number of failed units turned in, has probability mass function $\phi_j(k)$, $k \geqslant 1$, with finite mean. In the present context of batch inspection policy, the entire batch of failed units is repaired at the base with probability r_j, is shipped to the depot for repair with probability $(1 - r_j) \rho_j$, or is condemned with probability $(1 - r_j)(1 - \rho_j)$. Thus at the base, the requisitions are of three types: base-repairable, depot-repairable and condemnable.

Notation

λ_j = the rate of Poisson arrival of requisitions at location $j (j = 0, 1, \ldots, J)$,

ϕ_j = the probability mass function of order size at location $j (j = 0, 1, \ldots, J)$,

r_j = the probability that a batch is repaired at base $j (j = 1, 2, \ldots, J)$,

ρ_j = the probability that a batch which is not repairable at base j is repaired at the depot; $(1 - \rho_j)$ is, therefore, the probability of condemning a base-nonrepairable batch $(j = 1, 2, \ldots, J)$,

$(s_j - 1, s_j)$ = the procurement policy at base $j (j = 1, 2, \ldots, J)$,

(s_0, S_0) = the procurement policy at the depot

R_j = the deterministic repair time at location j $(j = 0, 1, \ldots, J)$,

τ_j = the deterministic delivery time from the depot to base $j (j = 1, 2, \ldots, J)$,

τ_0 = the deterministic procurement lead time from the external supplier to the depot,

$N_j(t)$ = the total number of requisitions that arrive at location j during the interval $(0, t]$ $(j = 0, 1, \ldots, J)$,

$N_j^B(t)$ = the number of requisitions at base j during the interval $(0, t]$ for which the entire batch was declared base repairable $(j = 1, 2, \ldots, J)$,

$N_j^D(t)$ = the number of requisitions at base j during the interval $(0, t]$ for which the entire batch was sent to the depot for repair $j = 1, 2, \ldots, J)$,

$N_j^C(t)$ = the number of requisitions at base j during the interval $(0, t]$ for which the entire batch was condemned $(j = 1, 2, \ldots, J)$,

$N_j^0(t)$ = the total number of requisitions placed at the depot by base j during the interval $(0, t]$ $(j = 1, 2, \ldots, J)$,

$N_0^C(t)$ = the total number of requisitions at the depot during the interval $(0, t]$ as a consequence of condemnations at the lower echelon,

$N_0^D(t)$ = the total number of requisitions at the depot during the interval $(0, t]$ for which the batches were found depot repairable.

The variables $D_j(t)$, $D_j^B(t)$, $D_j^D(t)$, $D_j^C(t)$, $D_j^0(t)$, $D_0^C(t)$ and $D_0^D(t)$ shall denote the number of units demanded by the corresponding requisitions $N_j(t)$, $N_j^B(t)$, $N_j^D(t)$, $N_j^C(t)$, $N_j^0(t)$, $N_0^C(t)$, and $N_0^D(t)$, respectively. For instance, $D_j^0(t)$ denotes the total number of units demanded from the depot by base j during the interval $(0, t]$, $(j = 1, 2, \ldots, J)$; and so on.

$Q_j(t)$ = the in-repair inventory at time t at location $j (j = 0, 1, \ldots, J)$,

$Z_j(t)$ = the inventory position at time t at location $j (j = 0, 1, \ldots, J)$,

$B_j(t)$ = the number of backorders at time t at location j $(j = 0, 1, \ldots, J)$ (Negative backorders indicate on-hand inventory),

$U_j(t)$ = the total number of units on order plus in repair at time t at location $j (j = 0, 1, \ldots, J)$,

$P[n|m] = \dfrac{e^{-m}(m)^n}{n!}$, $n = 0, 1, \ldots$ (Poisson distribution with mean m),

$CP[k|\lambda t, f] = \displaystyle\sum_{n=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^n}{n!} f^{(n)}(k)$, $k = 0, 1, \ldots$ (Compound Poisson distribution with parameter λt and compounding distribution f);

$f^n(\cdot)$ = n-fold convolution of density function $f(\cdot)$,

E_0 = $\{s_0 + 1, s_0 + 2, \ldots, S_0\}$, the state space for inventory position at depot,

$N(t_1, t_2)$ = $N(t_2) - N(t_1^+)$ for the process $\{N(t), t \geqslant 0\}$.

Lower case letters denote a particular realization of a random variable.

We note the following implications of our assumptions:

(a) Obviously, $N_j(t) = N_j^B(t) + N_j^C(t) + N_j^D(t)$, for all $t \geqslant 0$. It can be easily seen (see references [5], [8]) that the processes $\{N_j^B(t),\ t \geqslant 0\}$, $\{N_j^C(t),\ t \geqslant 0\}$, and $\{N_j^D(t),\ t \geqslant 0\}$ are mutually independent Poisson processes with parameters $\lambda_j^B = r_j\lambda_j$, $\lambda_j^C = (1 - r_j)(1 - \rho_j)\lambda_j$, and $\lambda_j^D = (1 - r_j)\rho_j\lambda_j$, respectively, for $j = 1, 2, \ldots, J$. Consequently, the demand processes $\{D_j^B(t),\ t \geqslant 0\}$, $\{D_j^C(t),\ t \geqslant 0\}$ and $\{D_j^D(t),\ t \geqslant 0\}$ are compound Poisson processes with parameters λ_j^B, λ_j^C and λ_j^D, respectively, and have a common compounding distribution $\phi_j(\cdot)$.

(b) Because the bases use an $(s - 1, s)$ policy, $N_j^0(t) = N_j^C(t) + N_j^D(t)$, for all $t \geqslant 0$ and $j = 1, 2, \ldots, J$. Consequently, $\{N_j^0(t),\ t \geqslant 0\}$ is a Poisson process with parameter $\lambda_j^0 = \lambda_j^C + \lambda_j^D$. Furthermore, because the bases operate independently, $\{N_0^C(t),\ t \geqslant 0\}$ and $\{N_0^D(t),\ t \geqslant 0\}$ are independent Poisson processes with parameters $\lambda_0^C = \sum_{j=1}^{J} \lambda_j^C$ and $\lambda_0^D = \sum_{j=1}^{J} \lambda_j^D$, respectively. Moreover, $N_0(t) = N_0^C(t) + N_0^D(t)$ for all $t \geqslant 0$, and thus $\{N_0(t),\ t \geqslant 0\}$ is a Poisson process with parameter $\lambda_0 = \sum_{j=1}^{J} \lambda_j^0$. Again it can be shown (see reference [5]) that the demand processes $\{D_0(t),\ t \geqslant 0\}$, $\{D_0^C(t),\ t \geqslant 0\}$ and $\{D_0^D(t),\ t \geqslant 0\}$ are compound Poisson processes with parameters λ_0, λ_0^C and λ_0^D; and compounding distributions $\phi_0(\cdot)$, $\phi_0^C(\cdot)$ and $\phi_0^D(\cdot)$, respectively; where

$$\phi_0(k) = \frac{1}{\lambda_0} \sum_{j=1}^{J} \lambda_j^0 \phi_j(k), \quad \phi_0^C(k) = \frac{1}{\lambda_0^C} \sum_{j=1}^{J} \lambda_j^C \phi_j(k)$$

and

$$\phi_0^D(k) = \frac{1}{\lambda_0^D} \sum_{j=1}^{J} \lambda_j^D \phi_j(k); \quad \text{for } k = 1, 2, \ldots .$$

(c) As a consequence of the $(s_j - 1, s_j)$ policy at base j, $Z_j(t) = s_j$ for all $t \geqslant 0$, or equivalently, $U_j(t) - B_j(t) = s_j$ for $t \geqslant 0$. Then for any $b \in \{-s_j, -s_j + 1, \ldots, 0, 1, \ldots\}$ and for any $t \geqslant \tau_0 + \tau_j$; the event $B_j(t) = b$ occurs if and only if $U_j(t) = s_j + b$.

(d) Because of the infinite number of repair facilities and constant repair times, the units in repair at base j at time t $(\geqslant R_j)$ will be due to the base repairable failures occuring only in $(t - R_j, t]$; that is $Q_j(t) = D_j^B(t - R_j, t)$. Thus for $t \geqslant R_j$, $Q_j(t)$ is a compound Poisson with parameter $\lambda_j^B R_j$ and compounding function $\phi_j(\cdot)$. Similarly, $Q_0(t)$ is a compound Poisson with parameter $\lambda_0^D R_0$ and compounding distribution $\phi_0^D(\cdot)$.

We first describe the basic approach for determining stationary distribution of system backorders in Section 3. In Section 4, the stationary distribution of backorders/on-hand inventory at bases is obtained. The results for in-repair inventory are obtained in Section 5, and those for depot backorder/on-hand inventory are given in Section 6. The special cases of complete recoverability, nonrecoverability and unit order size are discussed in Section 7. Section 8 indicates applicability and applications of the results derived in the present study.

3. BASIC APPROACH FOR STATIONARY DISTRIBUTION OF SYSTEM BACKORDERS

We emphasize that the system backorder means the backorders at the bases. Depot backorders are of interest only in so far as they affect base backorders. Stationary distribution of $\{Z_0(t),\ t \geqslant 0\}$, the inventory position at the depot, will be required to derive the expressions for stationary distribution of backorders at bases. The depot, for this purpose, can be treated as a single location system where arrive two types of demands, recoverable and nonrecoverable, and the corresponding processes $\{D_0^R(t),\ t \geqslant 0\}$ and $\{D_0^C(t),\ t \geqslant 0\}$ are independent compound Poisson. Further, the depot inventory position changes only at the arrival epochs of nonrecoverable demands from the bases. It remains unchanged at the arrival epochs of recoverable demands because in case of recoverable demands, on-hand inventory decreases and in-repair inventory increases by the same amount with no change in inventory position. Then stationary distribution of $\{Z_0(t),\ t \geqslant 0\}$ is given by (see Sahin [4], Shanker [5], Tijms [9]),

$$(1)\quad \lim_{t\to\infty} Pr\{Z_0(t) = k \mid Z_0(0) = i\} = \P_0(k) = \begin{cases} \dfrac{m(S_0 - k)}{1 + M(S_0 - s_0 - 1)} & s_0 + 1 \leqslant k \leqslant S_0 - 1 \\[2ex] \dfrac{1}{1 + M(S_0 - s_0 - 1)} & k = S_0 \end{cases}$$

where

$$m(1) = \phi_0^C(1);\ M(k) = \phi_0^C(k) + \sum_{q=1}^{k-1} \phi_0^C(k-q)\,m(q), \qquad k \geqslant 2$$

and

$$M(k) = \sum_{p=1}^{k} m(p), \qquad k \geqslant 1.$$

To find the stationary distribution $B_j^*(b) = \lim_{t\to\infty} Pr\{B_j(t) = b\}$ at base j, we first obtain $Pr\{B_j(t) = b\}$, or equivalently $Pr\{U_j(t) = s_j + b\}$ for $b \in \{-s_j, -s_j + 1, \ldots, 0, 1, \ldots\}$ and then evaluate $B_j^*(\cdot)$. We shall follow the approach suggested by Kruse and Kaplan [1].

Referring to Figure 2, the only units that can arrive at base j from the depot by time t are those on order by time t_3. This depends on the total assets available at the depot by time t_3, the total demand at the depot during the interval $(t_1, t_3]$, and the sequence of arrivals of requisitions at the depot from bases during the interval $(t_1, t_3]$. This is so because the units on order through time t_1 will have arrived at the depot from the external supplier by time t_3. The total assets available at the depot by time t_3 include the units on hand minus any backorders at time t_1, the units on order at time t_1, the units in repair at time t_1 and the units received for repair during the interval $(t_1, t_2]$. This equals $z_0(t_1) + d_0^R(t_1, t_2)$. Now the following two mutually exclusive situations are possible:

(A): The total depot demand during the interval $(t_1, t_2]$ does not exceed the total assets available by time t_3; that is,

$$d_0^R(t_1, t_2) + d_0^C(t_1, t_2) < z_0(t_1) + d_0^R(t_1, t_2)$$

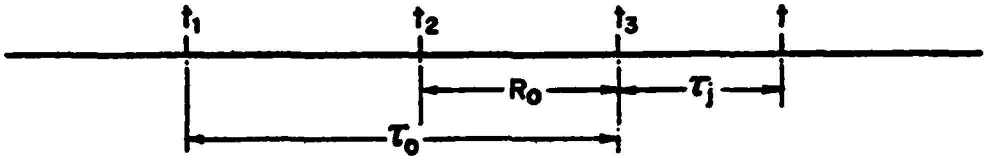


FIGURE 2. Time intervals at base j

or

$$d_0^C(t_1, t_2) < z_0(t_1).$$

Thus all the depot demands $d_0(t_1, t_2) = d_0^D(t_1, t_2) + d_0^C(t_1, t_2)$ are satisfied by time t_3. Only the depot demands $d_0(t_2, t_3)$ against the stock of $z_0(t_1) - d_0^C(t_1, t_2)$ units available by time t_3 determine how many demands could possibly remain unsatisfied by time t_3.

(B): The total demand during the interval $(t_1, t_2]$ exceeds the total available stock by time t_3; that is,

$$d_0^D(t_1, t_2) + d_0^C(t_1, t_2) > z_0(t) + d_0^D(t_1, t_2)$$

or

$$d_0^C(t_1, t_2) > z_0(t_1).$$

Thus there is no stock available at the depot at time t_2^+ to satisfy the demands $d_0(t_2, t_3)$. Also, there is no guarantee that all the $d_0^D(t_1, t_2)$ demands will be satisfied since this depends upon the sequence of arrivals of $N_0^C(t_1, t_2)$ and $N_0^D(t_1, t_2)$. Hence, the total demand $d_0(t_1, t_2)$ drawn against the amount of $z_0(t_1) + d_0^D(t_1, t_2)$ determines how many demands will remain unsatisfied by time t_3. We can write

$$\text{(2)} \qquad \begin{aligned} Pr\{B_j(t) = b\} &= Pr\{U_j(t) = s_j + b\} \\ &= \sum_{z_0(t_1) \in E_0} \sum_{d_0^C(t_1, t_2)}^{\infty} [Pr\{U_j(t) = s_j + b \mid D_0^C(t_1, t_2) \\ &= d_0^C(t_1, t_2);\ Z_0(t_1) = z_0(t_1)\} \cdot Pr\{D_0^C(t_1, t_2) \\ &= d_0^C(t_1, t_2);\ Z_0(t_1) = z_0(t_1)\}]. \end{aligned}$$

Using the independence of $D_0^C(t_1, t_2)$ and $Z_0(t_1)$ we can express Equation (2) corresponding to the cases (A) and (B) as follows:

$$\text{(3)} \qquad \begin{aligned} Pr\{U_j(t) = s_j + b\} = \sum_{z_0(t_1) \in E_0} &[Pr\{U_j(t) = s_j + b\}_A + Pr\{U_j(t) = s_j + b\}_B] \\ &\cdot Pr\{Z_0(t_1) = z_0(t_1)\} \end{aligned}$$

where

$$\text{(4)} \qquad \begin{aligned} Pr\{U_j(t) = s_j + b\}_A = \sum_{d_0^C(t_1, t_2) = 0}^{z_0(t_1)} &[Pr\{U_j(t) = s_j + b \mid D_0^C(t_1, t_2) = d_0^C(t_1, t_2); \\ &Z_0(t_1) = z_0(t_1)\} \cdot Pr\{D_0^C(t_1, t_2) = d_0^C(t_1, t_2)\}], \end{aligned}$$

and

$$(5)\qquad Pr\{U_j(t) = s_j + b\}_B = \sum_{d_0^C(t_1,t_2) > z_0(t_1)} [\{Pr\ U_j(t) = s_j + b | D_0^C(t_1, t_2) = d_0^C(t_1, t_2);\ Z_0(t_1) = z_0(t_1)\} \cdot Pr\{D_0^C(t_1, t_2) = d_0^C(t_1, t_2)\}].$$

For a given $z_0(t_1)$, Equations (4) and (5) represent $Pr\{B_j(t) = b\}$ for cases (A) and (B), respectively. In the next section we shall derive the expressions for $Pr\{U_j(t) = s_j + b\}_A$ and $Pr\{U_j(t) = s_j + b\}_B$ which upon substitution in Equation (3) will yield $Pr\{U_j(t) = s_j + b\}$.

4. STATIONARY DISTRIBUTION OF BACKORDERS/ON-HAND INVENTORY AT BASES

4.1 (A): $\mathbf{Pr\{U_j(t) = s_j + b\}_A}$

To evaluate $Pr\{U_j(t) = s_j + b\}_A$ using Equation (4), let

$$U_j(t) = U_j^1(t) + U_j^2(t),$$

where

$U_j^1(t)$ = the sum of units in repair at base j at time t and the units for which order were placed on the depot by base j during the interval $(t_3, t]$,

and

$U_j^2(t)$ = the units ordered from the depot by base j during the interval $(t_2, t_3]$ that remain unfilled by time t_3.

Because the arrival process is Poisson, $U_j^1(t)$ and $U_j^2(t)$ are independent. As mentioned earlier, all the demands levied on the depot from base j during the interval $(t_3, t]$ will remain on order at time t. Also the base repairable demands occurring only during the interval $(t - R_j, t]$ will be in the repair cycle at base j at time t. Thus $U_j^1(t) = D_j^B(t - R_j, t) + D_j^C(t_3, t) + D_j^D(t_3, t)$, and the probability distribution of $U_j^1(t)$ can be easily obtained. The probability distribution of $U_j^2(t)$ requires considering the sequence of arrivals of requisitions from the bases during the interval $(t_2, t_3]$. Once the probability distributions of $U_j^1(t)$ and $U_j^2(t)$ are obtained, Equation (4) can be evaluated by taking the convolution of $U_j^1(t)$ and $U_j^2(t)$. Since $U_j^1(t)$ is independent of $Z_0(t_1)$ and $D_0^C(t_1, t_2)$, we can write

$$(6)\quad Pr\{U_j(t) = s_j + b\}_A = \sum_{d_0^C(t_1,t_2)=0}^{z_0(t_1)} \left[\sum_{d=0}^{s_j+b} Pr\{U_j^1(t) = s_j + b - d\} \cdot Pr\{U_j^2(t) = d | D_0^C(t_1, t_2) = d_0^C(t_1, t_2);\ Z_0(t_1) = z_0(t_1)\}\right] \cdot Pr\{D_0^C(t_1, t_2) = d_0^C(t_1, t_2)\}.$$

As observed in Section 2, $U_j^1(t)$ has a compound Poisson distribution with parameter $\lambda_j^B R_j + (\lambda_j^C + \lambda_j^D)\tau_j$ and compounding distribution $\phi_j(\cdot)$. Therefore,

$$Pr\{U_j^1(t) = s_j + b - d\} = CP[s_j + b - d | \lambda_j^B R_j + (\lambda_j^C + \lambda_j^C)\tau_j, \phi_j]. \tag{7}$$

For the purpose of evaluating $Pr\{U_j^2(t) = d\}$, the demands at the depot can be viewed as arising from two sources. One, the base j for which the distribution is being determined and the other being set of the remaining bases (see Figure 3). Let us denote this set by σ, that is, $\sigma = \{1, 2, \dots, j-1, j+1, \dots, J\}$. Since the bases operate independently, the two sources are independent. For the source σ, we shall use the notations similar to those used for an individual base. Thus, $N_\sigma(t)$ denotes the total number of requisitions that arrive at source σ during the interval $(0, t]$, and so on. The processes $\{N_\sigma^B(t), t \geq 0\}$, $\{N_\sigma^C(t), t \geq 0\}$ and $\{N_\sigma^D(t), t \geq 0\}$ are mutually independent Poisson processes with parameters $\lambda_\sigma^B = \sum_{i \in \sigma} \lambda_i^B$, $\lambda_\sigma^C = \sum_{i \in \sigma} \lambda_i^C$ and $\lambda_\sigma^D = \sum_{i \in \sigma} \lambda_i^D$, respectively. Further, the demand processes $\{D_\sigma^B(t), t \geq 0\}$, $\{D_\sigma^C(t), t \geq 0\}$, and $\{D_\sigma^D(t), t \geq 0\}$ are compound Poisson processes with parameters λ_σ^B, λ_σ^C, and λ_σ^D, respectively. Their respective compounding distributions are:

$$\phi_\sigma^B(k) = \frac{1}{\lambda_\sigma^B} \sum_{i \in \sigma} \lambda_i^B \phi_i(k), \; \phi_\sigma^C(k) = \frac{1}{\lambda_\sigma^C} \sum_{i \in \sigma} \lambda_i^C \phi_i(k) \quad \text{and}$$

$$\phi_\sigma^D(k) = \frac{1}{\lambda_\sigma^D} \sum_{i \in \sigma} \lambda_i^D \phi_i(k); \quad \text{for } k \geq 1.$$

Also, the process $\{N_\sigma^0(t), t > 0\}$ is a Poisson process with parameter $\lambda_\sigma^0 = \sum_{i \in \sigma} \lambda_i^0$ and the demand process $\{D_\sigma^0(t), t \geq 0\}$ is a compound Poisson process with parameter λ_σ^0 and compounding distribution $\phi_\sigma^0(k) = \frac{1}{\lambda_\sigma^0} \sum_{i \in \sigma} \lambda_i^0 \phi_i(k)$, for $k \geq 1$.

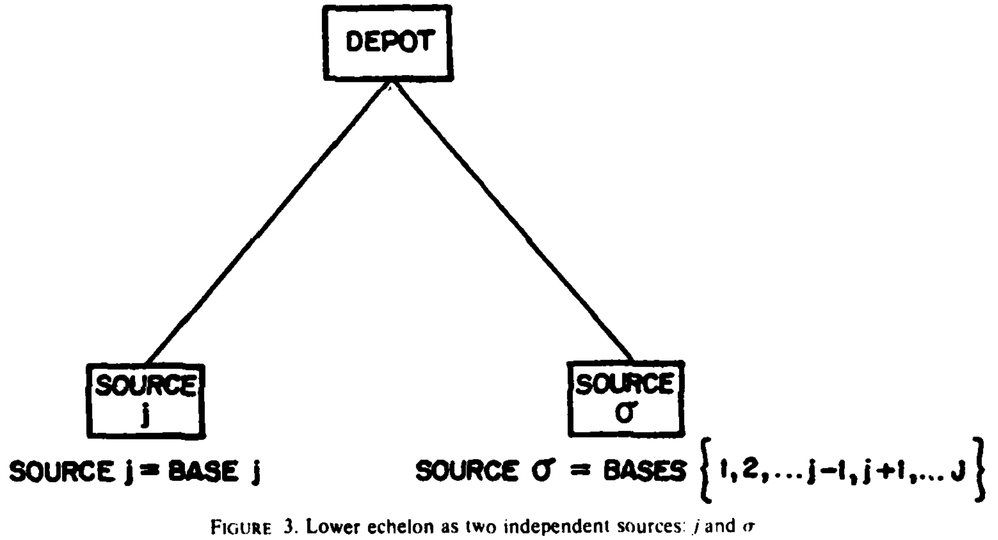


FIGURE 3. Lower echelon as two independent sources: j and σ

The random variable $U_j^2(t)$ represents the units ordered from the depot by base j during the interval $(t_2, t_3]$ that remain unfilled by time t_3. Referring to Equation (6), we shall obtain

$Pr\{U_j^2(t) = d | D_0^C(t_1, t_2) = d_0^C(t_1, t_2); Z_0(t_1) = z_0(t_1)\}$ by further conditioning on $N_i^0(t_2, t_3)$; $i = j, \sigma$. Let $RA = \{N_j^0(t_2, t_3) = n_j; N_\sigma^0(t_2, t_3) = n_\sigma; D_0^C(t_1, t_2) = d_0^C(t_1, t_2); Z_0(t_1) = z_0(t_1)\}$. Thus, we first obtain $Pr\{U_j^2(t) = d | RA\}$, and to do so we need to know the number of requisitions at the depot placed during $(t_2, t_3]$ and those completely satisfied by time t_3. Let

$N_i'(t_2, t_3)$ = the number of requisitions from source i that arrived during $(t_2, t_3]$ and are completely satisfied by time t_3; for $i = j, \sigma$.

Suppose we are given RA and $N_j'(t_2, t_3) = n_j'$. Then $U_j^2(t) = d$ if and only if the sum of the demands due to the unsatisfied $n_j - n_j'$ requisitions and unsatisfied units of possibly a partially satisfied requisition (if from base j) equals d. Let EX denote the number of units supplied to the requisition whose demand is only partially met. The range of the random variable EX is from 0 to $z_0(t_1) - d_0^C(t_1, t_2)$. When $EX = 0$, there is no partially satisfied requisition and when $EX = z_0(t_1) - d_0^C(t_1, t_2)$, no requisition is completely satisfied and $z_0(t_1) - d_0^C(t_1, t_2)$ units are supplied to the first requisition, if any (see Figure 4).

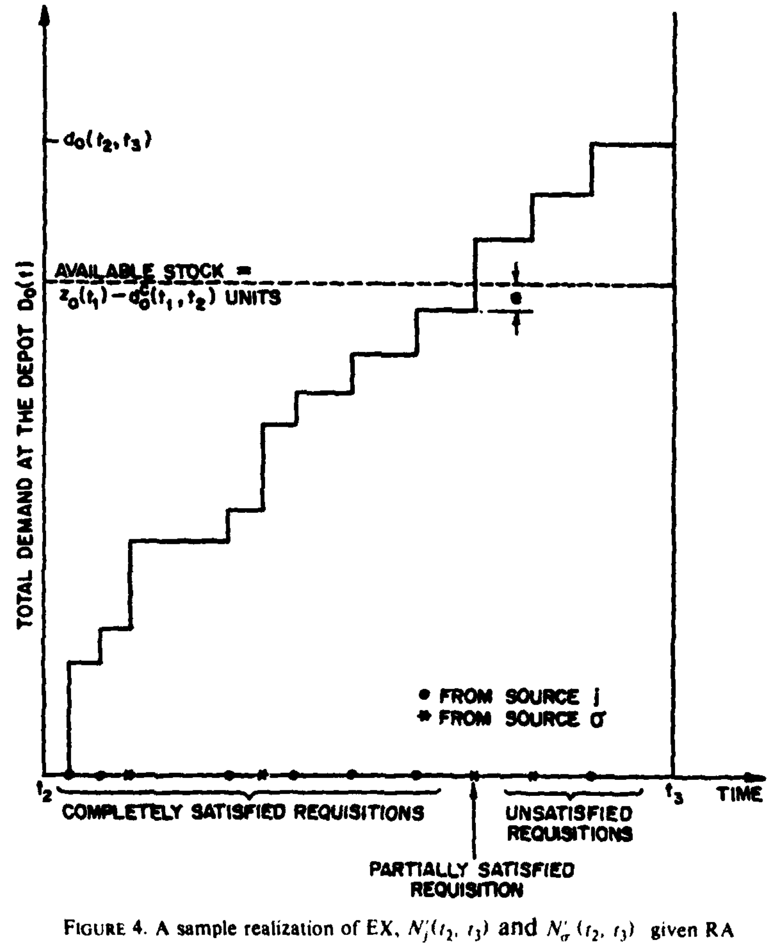


FIGURE 4. A sample realization of EX, $N_j'(t_2, t_3)$ and $N_\sigma'(t_2, t_3)$ given RA

Let us introduce an indicator variable I such that $I = i$ if the partially satisfied requisition is from source i; $i = j, \sigma$. Then we have the following:

$$\text{(8)} \quad Pr\{U_j^2(t) = d | RA;\ EX = 0,\ N_j'(t_2, t_3) = n_j';\ N_\sigma'(t_2, t_3) = n_\sigma'\} = \phi_j^{(n_j - n_j')}(d);$$

$$\text{(9)} \quad Pr\{U_j^2(t) = d | RA;\ EX = e > 0,\ N_j'(t_2, t_3) = n_j';\ N_\sigma'(t_2, t_3) = n_\sigma';\ I = \sigma\} = \phi_j^{(n_j - n_j')}(d);$$

and

$$\text{(10)} \quad Pr\{U_j^2(t) = d | RA;\ EX = e > 0,\ N_j'(t_2, t_3) = n_j';\ N_\sigma'(t_2, t_3) = n_\sigma';\ I = j\} = \sum_{k>0}^{d} \phi_j(k+e) \cdot \phi_j^{(n_j - n_j' - 1)}(d-k).$$

The $Pr\{U_j^2(t) = d | RA\}$ can be obtained from Equations (8-10) by first computing $Pr\{EX = e(>0);\ I = i,\ N_j'(t_2, t_3) = n_j';\ N_\sigma'(t_2, t_3) = n_\sigma' | RA\}$, for $i = j, \sigma$. Let Y_k denote the number of units demanded from the depot by the kth $(k \geqslant 1)$ requisition during $(t_2, t_3]$.

We consider the following two situations. One, where not all the requisitions during $(t_2, t_3]$ are satisfied by time t_3, and the other, where they are all satisfied.

(i): $0 \leqslant n_j' + n_\sigma' < n_j + n_\sigma$; $n_j + n_\sigma \geqslant 1$, not all the requisitions in $(t_2, t_3]$ are satisfied by time t_3. We wish to compute

$$\text{(11)} \quad Pr\{EX = e;\ I = j;\ N_j'(t_2, t_3) = n_j';\ N_\sigma'(t_2, t_3) = n_\sigma' | RA\} = Pr\{Y_1 + Y_2 + \ldots + Y_{n_j'+n_\sigma'} = z_0(t_1) - d_0^C(t_1, t_2) - e;$$

$Y_{n_j'+n_\sigma'+1} > e$; out of the first $(n_j' + n_\sigma'$ requisitions at the depot n_j' are from source j, and $(n_j' + n_\sigma' + 1)^{st}$ requisition is from source $j | RA\}$.

We have

$$\text{(12)} \quad Pr\{Y_1 + Y_2 + \ldots + Y_{n_j'+n_\sigma'} = z_0(t_1) - d_0^C(t_1, t_2) - e;\ Y_{n_j'+n_\sigma'+1} > e | RA\} = \phi_0^{(n_j'+n_\sigma'}(z_0(t_1) - d_0^C(t_1, t_2) - e) \cdot \sum_{p>e} \phi_j(p).$$

$Pr\{$out of the first $(n_j' + n_\sigma')$ requisitions at the depot, n_j' are from base j, and $(n_j' + n_\sigma' + 1)^{st}$ requisition is from source $j | RA\}$ can be derived using the results on sequences of Poisson arrivals (see Simon [8], Shanker [5]); and is given by

$$\text{(13)} \quad \frac{n_j' + 1}{n_j' + n_\sigma' + 1} \cdot \frac{\binom{n_j}{n_j'+1}\binom{n_\sigma}{n_\sigma'}}{\binom{n_j + n_\sigma}{n_j' + n_\sigma' + 1}}.$$

Substituting Equation (12), and Equation (13) upon simplification, into Equation (12) we have

(14) $Pr\{EX = e;\ I = j;\ N_j'(t_2,\ t_3) = n_j';\ N_\sigma'(t_2,\ t_3) = n_\sigma' | RA\}$

$$= \frac{(n_j - n_j')}{(n_j + n_\sigma) - (n' + n_\sigma')} \frac{\binom{n_j}{n_j'}\binom{n_\sigma}{n_\sigma'}}{\binom{n_j + n_\sigma}{n_j' + n_\sigma'}} \left\{\phi_0^{(n_j'+n')}(z_0(t_1) - d_0^C(t_1,\ t_2) - e) \cdot \sum_{p>e} \phi_j(p)\right\},$$

for $e = 0,\ 1,\ \ldots,\ z_0(t_1) - d_0^C(t_1,\ t_2) - (n_j' + n_\sigma');\ 0 \leqslant n_j' \leqslant n_j$.

Similarly,

(15) $Pr\{EX = e;\ I = \sigma;\ N_j'(t_2,\ t_3) = n_j';\ N_\sigma'(t_2,\ t_3) = n_\sigma' | RA\}$

$$= \frac{(n_\sigma - n_\sigma')}{(n_j + n_\sigma) - (n' + n_\sigma')} \frac{\binom{n_j}{n_j'}\binom{n_\sigma}{n_\sigma'}}{\binom{n_j + n_\sigma}{n_j' + n_\sigma'}} \left\{\phi_0^{(n_j'+n_\sigma')}(z_0(t_1) - d_0^C(t_1,\ t_2) - e) \cdot \sum_{p>e} \phi_\sigma^0(p)\right\},$$

for $e = 0,\ 1,\ \ldots,\ z_0(t_1) - d_0^C(t_1,\ t_2) - (n_j' + n_\sigma');\ 0 \leqslant n_\sigma' \leqslant n_\sigma$.

On summing Equations (14) and (15), we have

(16) $Pr\{EX = e;\ N_j'(t_2,\ t_3) = n_j';\ N_\sigma'(t_2,\ t_3) = n_\sigma' | RA\}$

$$= \frac{\binom{n_j}{n_j'}\binom{n_\sigma}{n_\sigma'}}{\binom{n_j + n_\sigma}{n_j' + n_\sigma'}} \phi_0^{(n_j'+n_\sigma')}(z_0(t_1) - d_0^C(t_1,\ t_2) - e) \cdot \sum_{p>e} \phi_0(p)$$

for $e = 0,\ 1,\ \ldots,\ z_0(t_1) - d_0^C(t_1,\ t_2) - (n_j' + n_\sigma')$, and $0 \leqslant n_j' + n_\sigma' < n_j + n_\sigma$

(ii): $n_j' + n_\sigma' = n_j + n_\sigma (\geqslant 0)$, all the requisitions in $(t_2,\ t_3]$ are satisfied by time t_3. In this case we have

(17) $Pr\{EX = e;\ N_j'(t_2,\ t_3) = n_j';\ N_\sigma'(t_2,\ t_3) = n_\sigma' | RA\}$

$$= Pr\{(Y_1 + Y_2 + \ldots + Y_{n_j'+n'}) \leqslant z_0(t_1) - d_0^C(t_1,\ t_2)\}$$

$$
= \left| \begin{array}{ll}
\displaystyle\sum_{p=n_j'+n_\sigma'}^{z_0(t_1)-d_0^C(t_1,t_2)} \phi_0^{(n_j'+n')}(p), & \text{for } e = 0,\ n_j' = n_j,\ n_\sigma' = n_\sigma, \\
 & \quad n_j + n_\sigma \geqslant 1; \\
1, & \text{for } e = z_0(t_1) - d_0^C(t_1,\ t_2), \\
 & \quad n_j' = n_j = n_\sigma' = n_\sigma = 0; \\
0, & \text{otherwise.}
\end{array} \right.
$$

From Equations (8-10) and Equations (11-17) we can obtain $Pr\{U_j^2(t) = d|RA\}$. For computational convenience, however, we consider the following two cases: one where all the demands from base j during $(t_2,\ t_3]$ are satisfied by time t_3, that is, $d = 0$; and the other where some demands from base j during $(t_2,\ t_3]$ remain unsatisfied by time t_3; that is, $d \geqslant 1$.

(a): $Pr\{U_j^2(t) = 0|RA\}$: From Equations (9) and (10) we conclude that given RA, $U_j^2(t) = 0$ if and only if $n_j' = n_j(\geqslant 0)$. For the case when all the requisitions at the depot in $(t_2,\ t_3]$ are satisfied by time t_3; that is, when $n_j' + n_\sigma' = n_j + n_\sigma$, $Pr\{U^2(t) = 0|RA\} = 1$. Then from Equation (17), we have

$$
Pr\{U_j^2(t) = 0|RA\} = \begin{cases} 1 & \text{for } n_j + n_\sigma = 0 \\ \displaystyle\sum_{p=n_j+n_\sigma}^{z_0(t_1)-d_0^C(t_1,t_2)} \phi_0^{(n_j+n_\sigma)}(p) & \text{for } n_j + n_\sigma \geqslant 1. \end{cases}
$$

In other words,

$$
(18) \qquad Pr\{U_j^2(t) = 0|RA\} = PRO1 \text{ (say)}
$$

$$
= \sum_{p=n_j+n_\sigma}^{z_0(t_1)-d_0^C(t_1,t_2)} \phi_0^{(n_j+n_\sigma)}(p).
$$

On the other hand, when $n_j' + n' < n_j + n_\sigma$; that is, when not all the requisitions at the depot during $(t_2,\ t_3]$ are satisfied by time t_3, then $n_j = n_j'$ implies that $n_\sigma' < n_\sigma$. From Equation (15), after simplification, we have

$$
(19) \quad Pr\{U_j^2(t) = 0|RA\} = PRO2 \text{ (say)}
$$

$$
= \sum_{n_\sigma'=0}^{n_\sigma-1} \left[\frac{n_\sigma!(n_j + n_\sigma')!}{n_\sigma'!(n_j + n_\sigma)!} \left\{ \sum_e \phi_0^{(n_j+n_\sigma')}(z_0(t_1) - d_0^C(t_1,\ t_2) - e) \cdot \sum_{p>e} \phi_0(p) \right\} \right],
$$

where the range of summation for e is from 0 to $z_0(t_1) - d_0^C(t_1,\ t_2) - n_j - n_\sigma'$.

(b): $Pr\{U_j^2(t) = d|RA\}$, $d \geqslant 1$: From Equations (9) and (10) it is clear that given RA, $U_j^2(t) = d(\geqslant 1)$ if $n_j' < n_j$ and thus $n_j' + n_\sigma < n_j + n_\sigma$ for $n_j \geqslant 1$. From Equations (8-10) and Equations (14-16), it follows that

$$(20)\quad Pr\{U_j^2(t) = d|RA\} = PRD \text{ (say)}$$

$$= \sum_{n_j'=0}^{n_j-1} \sum_{n_\sigma'=0}^{n_\sigma} \frac{\binom{n_j}{n_j'}\binom{n_\sigma}{n_\sigma'}}{\binom{n_j+n_\sigma}{n_j'+n_\sigma}} \sum_e \phi_0^{(n_j'+n_\sigma')}(z_0(t_1) - d_0^C(t_1,\ t_2) - e)$$

$$\cdot \Bigg\{ \Bigg[\frac{n_j - n_j'}{(n_j + n_\sigma) - (n_j' + n_\sigma')} \sum_{k=1}^{d} \phi_j(k+e) \cdot \phi_j^{(n_j - n_j' - 1)}(d-k) \Bigg]$$

$$+ \Bigg[\frac{n_\sigma - n_\sigma'}{(n_j + n_\sigma) - (n_j' + n_\sigma')} \phi_j^{(n_j' - n_\sigma')}(d) \sum_{p>e} \phi_\sigma^0(p) \Bigg] \Bigg\},$$

where the range for summation of e is from 0 to $z_0(t_1) - d_0^C(t_1,\ t_2) - n_j' - n_\sigma'$.

Now from Equations (18), (19) and (20), we obtain $Pr\{U_j^2(t) = d|D_0^C(t_1,\ t_2) = d_0^C(t_1,\ t_2);\ Z_0(t_1) = z_0(t_1)\}$ by enumerating over $N_j(t_2,\ t_3)$ and $N_\sigma(t_2,\ t_3)$. Substituting the resulting expression, and Equation (7) into Equation (6) we obtain $Pr\{U_j(t) = S_j + b\}_A$. Thus

$$(21)\quad Pr\{U_j(t) = s_j + b\}_A$$

$$= \sum_{d_0^C(t_1,t_2)=0}^{z_0(t_1)} \Bigg[CP[s_j + b|\lambda_j^B R_j + \lambda_j^0 \tau_j,\ \phi_j] \cdot \Bigg\{ \sum_{n_j+n_\sigma=0}^{z_0(t_1)-d_0^C(t_1,t_2)} (PRO1) \cdot p[n_j + n_\sigma|\lambda_0 R_0]$$

$$+ \Bigg\{ \sum_{n_j=0}^{z_0(t_1)-d_0^C(t_1,t_2)} \sum_{n_\sigma=1}^{\infty} (PRO2) \cdot P[n_\sigma|\lambda_\sigma^0 R_0] \cdot P[n_j|\lambda_j^0 R_0] \Bigg\} \Bigg\}$$

$$+ \sum_{d=1}^{s_j+b} CP[s_j + b - d|\lambda_j^B R_j + \lambda_j^0 \tau_j,\ \phi_j] \cdot \Bigg\{ \sum_{n_j=1}^{\infty} \sum_{n_\sigma=0}^{\infty} (PRD) \cdot P[n_\sigma|\lambda_\sigma^0 R_0] \cdot P|n_j|\lambda_j^0 R_0] \Bigg\} \Bigg]$$

$$\cdot CP[d_0^C(t_1,\ t_2)|\lambda_0^C(\tau_0 - R_0),\ \phi_0^C].$$

4.2 (B): $\mathbf{Pr}\{U_j(t) = s_j + b\}_B$.

To obtain the expression for $Pr\{U_j(t) = s_j + b\}_B$, we note that in case B all demands levied on the depot during interval $(t_2,\ t]$ remain unfilled by time t. In addition, some demands from the base j during the interval $(t_1,\ t_2]$ may remain unfilled by time t_3. Let for this case

$U_j^1(t)$ = the sum of units in repair at base j at time t and the units for which orders were placed on the depot by base j during $(t_2,\ t]$; that is, $U_j^1(t) = D_j^B(t - R_j,\ t) + D_j^C(t_2,\ t) + D_j^D(t_2,\ t)$;

and

$U_j^2(t)$ = the units ordered from the depot by base j during $(t_1,\ t_2]$ that are unfilled by time t_3.

Obviously $U_j(t) = U_j^1(t) + U_j^2(t)$, and $U_j^1(t)$ and $U_j^2(t)$ are independent. Therefore, the probability distribution of $U_j(t)$ can be obtained through the convolution of $U_j^1(t)$ and $U_j^2(t)$. The probability distribution of $U_j^1(t)$ can be readily obtained while that of $U_j^2(t)$ involves consideration of the sequence of arrivals of requisitions at the depot during $(t_1, t_2]$. Noting that $U_j^1(t)$ is independent of $Z_0(t_1)$ and $D_0^C(t_1, t_2)$ and that $Pr\{U_j^1(t) = k\} = CP[k|\lambda_j^B R_j + \lambda_j^0(\tau_j + r_0), \phi_j]$ from Equation (5) we have

$$(22) \qquad Pr\{U_j(t) = s_j + b\}_B$$

$$= \sum_{d_0^C(t_1,t_2) > z_0(t_1)}^{\infty} \left[\sum_{d=0}^{s_j+b} CP[s_j + b - d|\lambda_j^B R_j + \lambda_j^0(\tau_j + R_0), \phi_j]\right.$$

$$\left.\cdot Pr\{U_j^2(t) = d|D_0^C(t_1, t_2) = d_0^C(t_1, t_2);\ Z_0(t_1) = z_0(t_1)\}\right]$$

$$\cdot Pr\{D_0^C(t_1, t_2) = d_0^2(t_1, t_2)\}.$$

We shall obtain $Pr\{U_j^2(t) = d|D_0^C(t_1, t_2 = d_0^C(t_1, t_2);\ Z_0(t_1) = z_0(t_1)\}$ by further conditioning on $N_i^C(t_1, t_2)$, $N_i^D(t_1, t_2)$ and $D_i^D(t_1, t_2)$ for $i = j, \sigma$. We temporarily denote $d_0^C(t_1, t_2)$ by d_0^C. Let

$$RB1 = \{N_i^C(t_1, t_2) = n_i^C;\ N_i^D(t_1, t_2) = n_i^D, \text{ (for } i = j, \sigma);\ Z_0(t_1) = z_0(t_1)\},$$

$$RB2 = \left\{\sum_{i=j,\sigma} D_i^C(t_1, t_2) = D_0^C(t_1, t_2) = d_0^C;\ \sum_{i=j,\sigma} D_i^D(t_1, t_2) = d_0^D\right\}$$

and

$$RB = RB1\ U\ RB2$$

Obviously, then

$$(23) \qquad Pr\{RB2|RB1\} = \phi_0^{C^{(n_j^C + N_\sigma^C)}}(d_0^C) \cdot \phi_0^{D^{(n_j^D + n_\sigma^D)}}(d_0^D).$$

To obtain $Pr\{U_j^2(t)|RB\}$ we need to consider the number and type (depot repairable or condemned) of requisitions that were placed by the depot during $(t_1, t_2]$ from each source and are completely satisfied by time t_3. For $i = j, \sigma$, let

$N_i'^C(t_1, t_2)$ = the number of condemnation type requisitions that were placed at the depot during $(t_1, t_2]$ by source i and are completely satisfied by time t_3;

$N_i'^D(t_1, t_2)$ = the number of depot repairable type requisitions that were placed at the depot during $(t_1, t_2]$ by source i and are completely satisfied by time t_3

Now suppose we are given RB, $N_i'^C(t_1, t_2) = n_i'^C$ and $N_i'^D(t_1, t_2 = n_i'^D$ for $i = j, \sigma$. Then $U_j^2(t) = d$ if and only if the sum of the demands due to unsatisfied $(n_j^C + n_j^D - n_j'^C - n_j'^D)$ requisitions and unsatisfied units of a possibly partially satisfied requisition, if from source (base) j, equals d. Here in case B we note that $0 \leqslant n_j'^C + n_j'^D + n_\sigma'^C + n_\sigma'^D) < (n_j^C + n_j^D + n_\sigma^C + n_\sigma^D)$. Further, we introduce an indicator variable I to indicate the type and source of the partially satisfied requisition. For $i = j, \sigma$ let

$$I = \begin{cases} iC. & \text{if the partially satisfied requisition is of condemnation type and is from source } i \\ iD. & \text{if the partially satisfied requisition is of depot repairable type and is from source } i. \end{cases}$$

Also, let EX again denote the number of units supplied to the partially satisfied requisition. Then proceeding in the manner similar to that used in deriving Equations (8), (9) and (10), we get

$$(24) \quad Pr\{U_j^2(t) = d \mid RB;\ EX = 0;\ (N_i'^C(t_1, t_2) = n_i'^C;\ N_i'^D(t_1, t_2) = n_i'^D,\ i = j, \sigma)\} = \phi_j^{(n_j^C + n_j^D - n_j'^C - n_j'^D)}(d)$$

$$(25) \quad Pr\{U_j^2(t) = d \mid RB;\ EX = e > 0;\ (N_i'^C(t_1, t_2) = n_i'^C;\ N_i'^D(t_1, t_2) = n_i'^D,\ i = j, \sigma);\ I = \delta\} = \phi_j^{(n_j^C + n_j^D - n_j'^C - n_j'^D)}(d) \quad \text{for } \delta = \sigma C, \sigma D$$

and

$$(26) \quad Pr\{U_j^2(t) = d \mid RB;\ EX = e > 0;\ (N_i'^C(t_1, t_2) = n_i'^C;\ N_i'^D(t_1, t_2) = n_i'^D,\ i = j, \sigma);\ I = \delta\} = \sum_{k>0}^{d} \phi_j(k + e)\phi_j^{(n_j^C + n_j^D - n_j'^C - n_j'^D)}(d - k) \quad \text{for } \delta = jC, jD$$

To compute $Pr\{U_j^2(t) = d \mid RB\}$ we now need to obtain $Pr\{EX = e,\ N_i'^C(t_1, t_2) = n_i'^C,\ N_i'^D(t_1, t_2) = n_i'^D,\ i = j, \sigma);\ I = \delta \mid RB\}$. This will be done following the approach used in deriving Equations (14-17), and using the results on sequences of Poisson arrivals (see Shanker [5]). The probability of exactly $n_j'^C$, $n_j'^D$, $n_\sigma'^C$, $n_\sigma'^D$ requisitions out of n_j^C, n_j^D, n_σ^C, n_σ^D, respectively, being satisfied given that a total of $(n_j'^C + n_j'^D + n_\sigma'^C + n_\sigma'^D)$ out of $n_j^C + n_j^D + n_\sigma^C + n_\sigma^D)$ is satisfied, is given by

$$(27) \qquad PS = \frac{\begin{pmatrix} n_j^C \\ n_j'^C \end{pmatrix}\begin{pmatrix} n_j^D \\ n_j'^D \end{pmatrix}\begin{pmatrix} n_\sigma^C \\ n_\sigma'^C \end{pmatrix}\begin{pmatrix} n_\sigma^D \\ n_\sigma'^D \end{pmatrix}}{\begin{pmatrix} n_j^C + n_j^D + n_\sigma^C + n_\sigma^D \\ n_j'^C + n_j'^D + n_\sigma'^C + n_\sigma'^D \end{pmatrix}}.$$

Let $\{Y_1, Y_2, \ldots, Y_{n_i^C}\}$ be the sequence of the number of units demanded from the depot by the n_i^C requisitions from source i. Similarly, let $\{Y_1, Y_2, \ldots, Y_{n_i^D}\}$ be the sequence of the number of units demanded from the depot by the n_i^D requisitions from source $i (i = j, \sigma)$. Then

(28) $Pr\{EX = e;\ (N_i'^C(t_1, t_2) = n_i'^C;\ N_i'^D(t_1, t_2) = n_i'^D,\ i = j, \sigma);\ I = kC \mid RB\},\ k = j, \sigma$ $= Pr\{(Y_1 + Y_2 + \ldots + Y_{n_j'^C}) + (Y_1 + Y_2 + \ldots Y_{n_j'^D}) + (Y_1 + Y_2 + \ldots + Y_{n_\sigma'^C}) + (Y_1 + Y_2 + \ldots + Y_{n_\sigma'^D}) = z_0(t_1) + (Y_1 + Y_2 + \ldots + Y_{n_j^D}) + (Y_1 + Y_2 + \ldots Y_{n_\sigma^D}) - e;\ Y_{n_k'^C+1} > e$; out of the first $(n_j'^C + n_j'^D + n_\sigma'^C + n_\sigma'^D)$ requisitions at the depot, $(n_j'^C + n_j'^D)$ are from source j; the next requisition is of condemnation type and is from source $k \mid (Y_1 + Y_2 + \ldots + Y_{n_j^D}) + (Y_1 + Y_2 + \ldots + Y_{n_\sigma^D}) = d_0^D;\ (Y_1 + Y_2 + \ldots + Y_{n_j^C}) + (Y_1 + Y_2 + \ldots + Y_{n_\sigma^C} = d_0^C;\ RB1\}$

$$= \left[\frac{n_k^C - n_k'^C}{(n_j^C + n_j^D + n_\sigma^C + n_\sigma^D) - (n_j'^C + n_j'^D + n_\sigma'^C + n_\sigma'^D)}\right] \cdot PS \cdot \left\{ \sum_{k_1} \phi_0^{C^{(n_j'^C + n_\sigma'^C)}} (z_0(t_1) + k_1 - e) \right.$$

$$\cdot \left\{ \sum_{k_2} \phi_k^C(e + k_2) \cdot \phi_0^{C^{(n_j^C + n_\sigma^C - n_j'^C - n_\sigma'^C - 1)}} (d_0^C - z_0(t_1) - k_1 - k_2) \right\}$$

$$\left. \cdot \left\{ \phi_0^{D^{(n_j'^D + n_\sigma'^D)}}(k_1) \cdot \phi_0^{D^{(n_j^D + n_\sigma^D - n_j'^D - n_\sigma'^D)}}(d_0^D - k_1) \right\} \right\} \Big/ \left\{ \phi_0^{D^{(n_j^D + n_\sigma^D)}}(d_0^C) \cdot \phi_0^{C^{(n_j^C + n_\sigma^C)}}(d_0^C) \right\}$$

where

$$n_k'^D \leqslant n_k^D (\geqslant 0),\ n_k'^C < n_k^C (\geqslant 1); \quad k = j, \sigma$$

$$e = 0, 1, \ldots, (z_0(t_1) + d_0^D - (n_j'^C + n_j'^D + n_\sigma'^C + n_\sigma'^D)),$$

and the ranges of k_1 and k_2 are

$$\max\{n_j'^D + n_\sigma'^D,\ n_j'^C + n_\sigma'^C - z_0(t_1) + e\} \leqslant k_1 \leqslant d_0^D - (n_j^D + n_\sigma^D - n_j'^D - n_\sigma'^D),$$

$$1 \leqslant k_2 \leqslant \min\{d_0^C - z_0(t_1) - k_1 - (n_j^C + n_\sigma^C - n_j'^C - n_\sigma'^C),\ d_0^C - e - (n_j^C + n_\sigma^C - n_j'^C - n_\sigma'^C)\}.$$

Similarly the probability distribution when the partially satisfied requisition is depot repairable type, is given by

(29) $Pr\{EX = e; (N_i'^C(t_1, t_2) = n_i'^C;\ N_i'^D(t_1, t_2) = n_i'^D;\ i = j, \sigma); I = kD | RB\},\ k = j, \sigma$

$$= \left[\frac{n_k^D - n_k'^D}{(n_j^C + n_j^D + n_\sigma^C + n_\sigma^D) - (n_j'^C + n_j'^D + n_\sigma'^C + n_\sigma'^D)}\right] \cdot PS$$

$$\left\{ \sum_{k_1} \phi_0^{C^{(n_j'^C + n_\sigma'^C)}} (z_0(t_1) + k_1 - e) \left\{ \sum_{k_2} \phi_k^D(e + k_2) \cdot \phi_0^{D^{(n_j^D + n_\sigma^D - n_j'^D - n_\sigma'^D - 1)}} (d_0^D - k_1 - k_2 - e) \right\} \right.$$

$$\left. \cdot \phi_0^{D^{(n_j'^D + n_\sigma'^D)}}(k_1) \cdot \phi_0^{C^{(n_j^C + n_\sigma^C - n_j'^C - n_\sigma'^C)}}(d_0^C - z_0(t_1) - k_1 + e) \right\} \Big/ \left\{ \phi_0^{D^{(n_j^D + n_\sigma^D)}}(d_0^D) \cdot \phi_0^{C^{(n_j^C + n_\sigma^C)}}(d_0^C) \right\}$$

where

$$n_k'^C \leqslant n_k^C (\geqslant 0),\ n_k'^D < n_k^D (\geqslant 1); \quad k = j, \sigma; \ (\phi_j^D = \phi_j)$$

$$e = 0, 1, \ldots, (z_0(t_1) + d_0^D - (n_j'^C + n_j'^D + n_\sigma'^C + n_\sigma'^D)),$$

and the ranges of k_1 and k_2 are

$$\max\{n_j'^D + n_\sigma'^D,\ n_j'^C + n_\sigma'^C - z_0(t_1) + e\} \leqslant k_1 \leqslant d_0^C - z_0(t_1) + e - (n_j^C + n_\sigma^C - n_j'^C - n_\sigma'^C),$$

$$1 \leqslant k_2 \leqslant d_0^D - k_1 - e - (n_j^D + n_\sigma^D - n_j'^D - n_\sigma'^D).$$

The probability distribution of $U_j^2(t)$ can now be obtained from Equations (23-29). Letting $PS1 = (n_j^C + n_j^D + n_\sigma^C + n_\sigma^D) - (n_j'^C + n_j'^D + n_\sigma'^C + n_\sigma'^D)$ and after simplifications we have from Equations (22),

$$(30)\quad Pr\{U_j(t) = s_j + b\}_B = \sum_{d_0^C > z_0(t_1)} \sum_{d=0}^{s_j+b} CP[s_j + b - d | \lambda_j^B R_j + \lambda_j^0(\tau_j + R_0),\ \phi_j]$$

$$\cdot \left\{ \sum_{n_\sigma^D} \sum_{n_\sigma^C} \sum_{n_j^D} \sum_{n_j^C} \left\{ \sum_{n'^D_\sigma} \sum_{n'^C_\sigma} \sum_{n'^D_j} \sum_{n'^C_j} \left[\frac{PS}{PS1} \right] \sum_{d_0^D} \sum_{e} \{(n_\sigma^C - n'^C_\sigma)\, US(\sigma,\ C) \right.\right.$$

$$\left. + (n_\sigma^D - n'^D_\sigma)\, US(\sigma,\ D) + (n_j^C - n'^C_j)\, US(j,\ C) + (n_j^D - n'^D_j)\, US(j,\ D)\} \right\}$$

$$\cdot P[n_j^C | \lambda_j^C(\tau_0 - R_0)] \cdot P[n_j^D | \lambda_j^D(\tau_0 - R_0)] \cdot P[n_\sigma^C | \lambda_\sigma^C(\tau_0 - R_0)]$$

$$\left. \cdot P[n^D | \lambda_\sigma^D(\tau_0 - R_0)] \right\}$$

where

$$US(\sigma,\ C) = \sum_{k_1} \left\{ \phi_0^{C^{(n'^C_j + n'^C_\sigma)}}(z_0(t_1) + k_1 - e) \cdot \phi_0^{D^{(n'^D_j + n'^D_\sigma)}}(k_1) \left\{ \sum_{k_2} \phi_\sigma^C(e + k_2) \right.\right.$$

$$\cdot \left\{ \sum_{k_3=0}^{d} \phi_j^{(n_j^C - n'^C_j)}(k_3) \cdot \phi_\sigma^{C^{(n_\sigma^C - n'^C_\sigma - 1)}}(d_0^C - z_0(t_1) - k_1 - k_2 - k_3) \right.$$

$$\left.\left.\left. \cdot \phi_j^{(n_j^D - n'^D_j)}(d - k_3) \cdot \phi_\sigma^{D^{(n_\sigma^D - n'^D_\sigma)}}(d_0^D - k_1 - d + k_3) \right\}\right\}\right\};$$

$$US(\sigma,\ D) = \sum_{k_1} \left\{ \phi_0^{C^{(n'^C_j + n'^C_\sigma)}}(z_0(t_1) + k_1 - e) \cdot \phi_0^{D^{(n'^D_j + n'^D_\sigma)}}(k_1) \left\{ \sum_{k_2} \phi_\sigma^D(e + k_2) \right.\right.$$

$$\cdot \left\{ \sum_{k_3=0}^{d} \phi_j^{(n_j^C - n'^C_j)}(k_3) \cdot \phi_\sigma^{C^{(n_\sigma^C - n'^C_\sigma)}}(d_0^C - z_0(t_1) - k_1 - k_3 + e) \right.$$

$$\left.\left.\left. \cdot \phi_j^{(n_j^D - n'^D_j)}(d - k_3) \cdot \phi_\sigma^{D^{(n_\sigma^D - n'^D_\sigma - 1)}}(d_0^D - k_1 - k_2 + e - d + k_3) \right\}\right\}\right\};$$

$$US(j,\ C) = \sum_{k_1} \left\{ \phi_0^{C^{(n'^C_j + n'^C_\sigma)}}(z_0(t_1) + k_1 - e) \cdot \phi_0^{D^{(n'^D_j + n'^D_\sigma)}}(k_1) \left\{ \sum_{k_2=0}^{d} \phi_j(e + k_2) \right.\right.$$

$$\cdot \left\{ \sum_{k_3} \phi_j^{(n_j^C - n'^C_j - 1)}(k_3) \cdot \phi_\sigma^{C^{(n_\sigma^C - n'^C_\sigma)}}(d_0^C - z_0(t_1) - k_1 - k_2 - k_3) \right.$$

$$\cdot \phi_j^{(n_j^D - n_j'^D)}(d - k_2 - k_3) \cdot \phi_\sigma^{D^{(n_\sigma^D - n_\sigma'^D)}}(d_0^D - k_1 - d + k_2 + k_3)\Big\}\Big\}\Big\};$$

and

$$US(j,\ D) = \sum_{k_1} \Bigg\{ \phi_0^{C^{(n_j'^C + n_\sigma'^C)}}(z_0(t_1) + k_1 - e) \cdot \phi_0^{D^{(n_j'^D + n_\sigma'^D)}}(k_1) \Bigg\{ \sum_{k_2 = 0}^{d} \phi_j(e + k_2)$$

$$\Bigg\{ \sum_{k_3} \phi_j^{(n_j^C - n_j'^C)}(k_3) \cdot \phi_\sigma^{C^{(n_\sigma^C - n_\sigma'^C)}}(d_0^C - z_0(t_1) - k_1 - k_3 + e)$$

$$\cdot \phi_j^{(n_j^D - n_j'^D - 1)}(d - k_3) \cdot \phi_\sigma^{D^{(n_\sigma^D - n_\sigma'^D)}}(d_0^D - k_1 - k_2 + e - d + k_2 + k_3)\Big\}\Big\}\Big\};$$

The ranges of k_1 for enumeration in $US(\sigma,\ C)$ and $US(j,\ C)$ are as given in Equation (28), and those in $US(\sigma,\ D)$ and $US(j,\ D)$ are given in Equation (29). The ranges k_2 in $US(\sigma,\ C)$ and $US(\sigma,\ D)$ are also given in Equation (28) and (29), respectively. The ranges of k_3 in the $US(j,\ C)$ and $US(j,\ D)$ can similarly be obtained to ensure that $\phi^{(0)}(0) = 1$ and $\phi^{(m)}(n) = 0$ for $m < n$. The ranges of $n_\sigma'^D$, $n_\sigma'^C$, $n_j'^D$and $n_j'^C$are similarly taken to ensure that the combinatorial terms in PS are nonnegative.

Since the expressions for $Pr\{U_j(t) = s_j + b\}_A$ and $Pr\{U_j(t) = s_j + b\}_B$ given by Equations (21) and (30), respectively, are independent of t, we obtain the stationary distribution $\lim_{t \to \infty} Pr\{B_j(t) = b\} = B_j^*(b)$ by substituting these equations and Equation (1) in Equation (3).

5. STATIC^NARY DISTRIBUTION OF IN-REPAIR INVENTORY

As observed in Section 2, we note that the stationary distribution of the number of units in repair at base j is given from the results of a $M^{[\phi_j]}/R_j/\infty$ queueing system. As mentioned by Sherbrooke [7], the distribution is a compound Poisson with parameter $\lambda_j^B R_j$ and compounding distribution $\phi_j(\cdot)$, [using Palm's theorem]; that is,

$$(31) \qquad \lim_{t \to \infty} Pr\{Q_j(t) = k\} = \sum_{n=0}^{k} \phi_j^{(n)}(k) \frac{e^{-\lambda_j^B R_j}(\lambda_j^B R_j)^n}{n!}, \quad \begin{array}{l} k = 0,\ 1,\ \ldots; \\ j = 1,\ 2,\ \ldots J. \end{array}$$

Similarly, the stationary distribution of the number of units at the depot is given by,

$$(32) \qquad \lim_{t \to \infty} Pr\{Q_0(t) = k\} = \sum_{n=0}^{k} \phi_0^{D^{(n)}}(k) \frac{e^{-\lambda_0^P R_0}(\lambda_0^P R_0)^n}{n!}, \quad k = 0,\ 1,\ \ldots$$

6. STATIONARY DISTRIBUTION OF DEPOT BACKORDERS/ON-HAND INVENTORY

As mentioned earlier, the depot can be treated as a single location which receives recoverable and nonrecoverable types of demand generated by independent compound Poisson processes $\{D_0^D(t),\ t \geqslant 0\}$ and $\{D_0^C(t),\ t \geqslant 0\}$, respectively. Let $X_0(t)$ represent the depot inventory level at time t which consists of the units on-hand minus any backorders. The positive values of $X_0(t)$ indicate on-hand inventory while the negative values indicate backorders $B_0(t)$ at time t, that is,

$$B_0(t) = \max(0, -X_0(t)).$$

It is more convenient here to obtain the stationary distribution in terms of $X_0(t)$ than for $B_0(t)$ directly. Let $E_x = \{S_0, S_0 - 1, \ldots 1, 0, -1, -2, \ldots\}$ denote the state space of $X_0(t)$.

Now, any thing on order from an external supplier at time $t - \tau_0$ will have arrived by time t, thus the number of units received via procurement during $(t - \tau_0, t]$ is the number of units on order at time $t - \tau_0$. The number of units arriving from the repair shop during $(t - \tau_0, t]$ is the number of units repaired during this interval, which equals $Q_0(t - \tau_0)$ $D_0^D(t = \tau_0, t) - Q_0(t)$. Therefore,

$$\begin{aligned} X_0(t) &= X_0(t - \tau_0) + (\text{units on order at time } t - \tau_0) + \text{units repaired during} \\ &\quad (t - \tau_0, t] - (\text{total demand during } (t - \tau_0, t]) \\ &= X_0(t - \tau_0) + (\text{units on order at time } t - \tau_0) + Q_0(t - \tau_0) + D_0^D(t - \tau_0, t) \\ &\quad - Q_0(t) - D_0^D(t - \tau_0, t) - D_0^C(t - \tau_0, t) \\ &= Z_0(t - \tau_0) - Q_0(t) - D_0^C(t - \tau_0, t). \end{aligned}$$

Then for all $i \in E_0$ and $j \in E_x$,

$$\begin{aligned} (33)\quad & Pr\{X_0(t) = j \mid Z_0(0) = i\} \\ &= \sum_{k \in E_0} Pr\{X_0(t) = j \mid Z_0(t - \tau_0) = k,\ Z_0(0) = i\} \cdot Pr\{Z_0(t - \tau_0) = k \mid Z(0) = i\} \\ &= \sum_{k \in E_0} Pr\{Z_0(t - \tau_0) - Q_0(t) - D_0^C(t - \tau_0, t) = j \mid Z_0(t - \tau_0) = k,\ Z_0(0) = i\} \\ &\quad \cdot Pr\{Z_0(t - \tau_0) = k \mid Z_0(0) = i\} \\ &= \sum_{k \in E_0} \Bigg[\sum_{m=0}^{\infty} Pr\{Q_0(t) + D_0^C(t - \tau_0, t) = k - j \mid Q_0(t - \tau_0) = m; Z_0(t - \tau_0) = k; \\ &\qquad Z_0(0) = i\} \cdot Pr\{Q_0(t - \tau_0) = m \mid Z_0(t - \tau_0) = k; Z_0(0) = i\} \Bigg] \\ &\quad \cdot Pr\{Z_0(t - \tau_0) = k \mid Z_0(0) = i\}. \end{aligned}$$

It can be seen that $Q_0(t)$ and $Z_0(t)$ are independent for any $t > 0$ (see for proof reference [5]). Then upon simplification we get

$$\begin{aligned} (34)\quad & Pr\{X(t) = j \mid Z_0(t - \tau_0) = k,\ Z_0(0) = i\} = Pr\{D_0^C(t - \tau_0, t) + Q_0(t) = k - j \mid Z_0(0) = i\} \\ &= Pr\{D_0^C(t - \tau_0, t) + Q_0(t) = k - j \mid Z_0(0) = i\} \\ &= \sum_{d=0}^{k-j} Pr\{Q_0(t) = k - j - d \mid Z_0(0) = i\} \cdot Pr\{D_0^C(t - \tau_0, t) = d\}. \end{aligned}$$

Substituting Equation (34) into Equation (33) and after taking the limit as $t \to \infty$, we get

$$\begin{aligned} X_0^*(j) &= \lim_{t\to\infty} Pr\{X_0(t) = j \mid Z_0(0) = i\} \\ &= \sum_{k\in E_0} \left[\sum_{d=0}^{k-j} \lim_{t\to\infty} Pr\{Q_0(t) = k - j - d \mid Z_0(0) = i\} \right. \\ &\quad \left. \cdot \lim_{t\to\infty} Pr\{D_0^C(t-\tau_0,\ t) = d\} \right] \cdot \pi_0(k). \end{aligned}$$

From Equation (32) then we have

$$(35) \qquad X_0^*(j) = \sum_{k\in E_0} \left\{ \sum_{d=0}^{k-j} CP[k - j - d \mid \lambda_0^D R_0,\ \phi_0^D] \cdot CP[d \mid \lambda_0^C \tau_0,\ \phi_0^C] \right\} \cdot \pi_0(k)$$

$$\text{for } j = S_0,\ S_0 - 1,\ \ldots 1,\ 0,\ -1,\ -2,\ \ldots$$

or in terms of the backorders

$$(36) \qquad B_0^*(b) = \sum_{k\in E_0} \left\{ \sum_{d=0}^{k+b} CP[k + b - d \mid \lambda_0^D R_0,\ \phi_0^D] \cdot CP[d \mid \lambda_0^C \tau_0,\ \phi_0^C] \right\} \cdot \pi_0(k)$$

$$\text{for } b = -S_0,\ -S_0 + 1,\ \ldots 0,\ 1,\ \ldots$$

7. SPECIAL CASES

The results on stationary distributions for special cases of complete recovery ($\rho_j = 1$, $j = 1, 2, \ldots, J$) and nonrecoverability ($\rho_j = 0$, $r_j = 0$; $j = 1, 2, \ldots, J$) can be derived from the expressions obtained in Sections (4-6). For complete recovery when there are no condemnations and the system, as referred to by Sherbrooke, is 'conservative', no procurement is made from the external supplier. The inventory position at the depot remains at a constant level S_0(say); that is, $Z_0(t) = S_0$ for all $t \geqslant 0$. Since $D_0^C(t) = 0$ for all $t \geqslant 0$, the Case (B) discussed in Section 4.2 will not arise and consequently Equation (3) reduces to $Pr\{U_j(t) = s_j + b\} = Pr\{U_j(t) = s_j + b\}_A$ with $z_0(t_1) = S_0$. Upon substituting $\lambda_0^C = 0$, and $z_0(t) = S_0$ in Equation (21) we can obtain the stationary distribution $B_j^*(b)$. The stationary distributions of in-repair inventory are given by Equations (31) and (32). Equations (35) and (36) similarly can be appropriately modified to yield stationary distribution of on-hand inventory and backorders, respectively.

For the classical case of nonrecoverability, that is, when the item is consumable, repair loop is absent at each location in the system, and $Q_j(t) = 0$ for $t \geqslant 0$ and $j = 0, 1, \ldots J$. The stationary distribution of $\{Z_0(t),\ t \geqslant 0\}$ can be obtained by noting that inventory position at the depot now changes at all arrival epochs of demands at the bases. $\{D_0(t),\ t \geqslant 0\}$ is a compound Poisson process with parameter $\lambda_0 = \sum_{j=1}^{J} \lambda_j$, and compounding distribution $\phi_0(\cdot) = \frac{1}{\lambda_0} \cdot \sum_{j=1}^{J} \lambda_j \phi_j(\cdot)$. Equation (1) can then be modified to give the stationary distribution in this case by replacing $\phi_0^C(\cdot)$ with $\phi_0(\cdot)$. For determining $B_j^*(b)$, we note that the cases (A) and (B) become $d_0(t_1,\ t_3) \leqslant z_0(t_1)$ and $d_0(t_1,\ t_3) > z_0(t_1)$, respectively. Equations (21) and (30) can be modified to give $B_j^*(b)$. Similarly, $X_0^*(b)$ can be obtained from Equation (36) as

$$X_0^*(j) = \sum_{k \in E_0} \left\{ \sum_{d=0}^{k-j} CP[d|\lambda_0^C\tau_0, \phi_0^C] \cdot \pi_0(k) \right\}, \tag{37}$$

a result quite familiar in classical consumable item inventory systems.

The results for the case of unit demand, that is, $\phi_j(1) = 1$, $j = 1, 2, \ldots, J$ can be similarly derived by noting that $\phi^m(n) = 1$ for $m = n$ and $\phi^m(n) = 0$ for $m \neq n$, and that $D_j(t)$, $D_j^B(t)$, $D_j^C(t)$, $D_j^D(t)$ are identical to $N_j(t)$, $N_j^B(t)$, $N_j^C(t)$, $N_j^D(t)$, respectively are are simple Poisson. Consequently, $D_0(t)$, $D_0^B(t)$, $D_0^C(t)$, $D_0^D(t)$ are identical to $N_0(t)$, $N_0^B(t)$, $N_0^C(t)$, $N_0^D(t)$, respectively, and are simple Poisson. The depot inventory position $Z_0(t)$ is then uniformly distributed over E_0. Equations (21) and (30) can be modified easily to yield $B_i^*(b)$, and it can be seen that the resulting expressions are the same as obtained by Simon [8]. For the unit demand case, the cases of complete recovery and nonrecoverability can be dealt in a manner similar to the one discussed above. For the case of complete recovery, the results are the same as given by Sherbrooke [7] for unit demand at the bases.

The results for the special cases described above are obtained by suitable modifications of the expressions derived in Sections (4-6). The details, however, can be found in references [5] and [8].

8. CONCLUSIONS

The exact expressions for stationary distributions of the depot inventory position, and of the number of backorders, the on-hand inventory, and the in-repair inventory at each location have been derived under the conditions of deterministic repair and lead times. From these expressions we can determine long-run average to formulate objective function of the total expected cost and can express the system performance measures such as service rate, ready rate etc. for the purpose of system optimization. The expressions obviously are computationally complex as they involve calculations of several combinatorial expressions and convolutions.

The results can also be used to assess the degree to which some approximate but computationally simpler models can serve as an approximation to our exact results. For example, for a unit demand case in a two-echelon conservative system, a study (reference [6]) on comparison of the exact results with those given by Sherbrooke's METRIC model which is approximate but much simpler to use, reveals that a considerable discrepancy exists between the two results especially when the depot spare stock level is low or when a major portion of the repairs is carried out at the depot. For the problem of optimal allocation of units of a spare item, it has been concluded in reference [6] that the METRIC model will suggest larger stocks at the bases with poor repair capability, than given by our exact results. Similar comparisons can be made for the problem of allocation of units in a multi-item system.

The results of the present analysis also apply to the situation where the three types of system demands (base-repairable, depot-repairable, condemnable) at the bases arrive independently in a Poisson manner from different sources.

ACKNOWLEDGMENT

The author acknowledges the valuable comments and suggestions made by the referee of this paper.

REFERENCES

[1] Kruse, W.K. and A.J. Kaplan, "On a Paper by Simon," Operations Research, *21*, 1318-1322 (1973).

[2] Muckstadt, J.A., "On the Probability Distribution for Inventory Position in Two-Echelon Continuous Review Systems," Technical Report No. 336, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, June 1977.

[3] Muckstadt, J.A., "Analysis of a Two-Echelon Inventory System in which all Locations Follow Continuous Review (s, S) Policies," Technical Report No. 337, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, September 1977.

[4] Sahin, I., "On the Stationary Analysis of Continuous Review (s, S) Inventory System with Constant Lead Times," Operations Research, *27*(4), 717-729 (1979).

[5] Shanker, K., "An Analysis of a Two-Echelon Inventory System for Recoverable Items," Unpublished Doctoral Dissertation, (also Technical Report No. 341), School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, June 1977.

[6] Shanker, K., "A Comparison of EXACT and METRIC Models for a Two-Echelon Conservative Inventory System for Recoverable Items—The Case of Unit Order Size," presented at the XXIV International Meeting, TIMS, June 1979, Hawaii.

[7] Sherbrooke, C.C., "METRIC: A Multi-Echelon Technique for Recoverable Item Control," Operations Research, *16*, 122-141 (1968).

[8] Simon, R.M., "Stationary Properties of a Two-Echelon Inventory Model for Low-Demand Items," Operations Research, *19*, 761-773 (1971).

[9] Tijms, H.C., *Analysis of (s, S) Inventory Models* (Mathematisch Centrum, Amsterdam, 1972).

ON A SHAPE ESTIMATOR OF WEISS

Ishay Weissman

Faculty of Industrial Engineering
and Management
Technion-Israel Institute of Technology
Haifa, Israel

ABSTRACT

The problem of estimating the shape-parameter of a distribution is considered. We introduce a class of estimators the distributions of which are independent of location and scale. An estimator proposed by Weiss [1] is a member of this class. We find the asymptotically most efficient estimator in this class which differs from that proposed by Weiss.

1. INTRODUCTION

Let $X_{n:1} \leqslant X_{n:2} \leqslant \dots \leqslant X_{n:n}$ be the order statistics from a sample of size n from a distribution function F with density $f(x)$. We assume that

$$(1.1) \qquad \begin{aligned} &f(x) = 0 \text{ for } x < \theta \\ &f(x) = c(x-\theta)^{\gamma-1}(1 + r(x-\theta)) \text{ for } x \geqslant \theta, \end{aligned}$$

where $c, \theta, \gamma, r(y)$ are all unknown except that we know $c > 0, \gamma > 0$ and $|r(y)| \leqslant Ky^{\beta}$ for all y in some interval $[0, \Delta]$, where $K, \beta, \Delta > 0$, again unknown. Suppose we want to estimate γ on the basis of $X_{n:1}, \dots, X_{n:k(n)}$, where $k(n) \to \infty$ as $n \to \infty$ and $k(n)/n \to 0$. Weiss [1] suggests the following simple estimator for the shape-parameter $\delta \equiv \gamma^{-1}$

$$(1.2) \qquad \hat{\delta} = \log \frac{X_{n:k(n)} - X_{n:1}}{X_{n:[k(n)/2]} - X_{n:1}} / \log 2,$$

which is consistent for δ but is neither asymptotically efficient, nor is it optimal in any ordinary sense. The purpose of this note is to suggest a similar estimator which is more efficient and is optimal in some sense.

2. THE OPTIMAL ESTIMATOR

It is shown in Weiss [1] that under (1.1), for all asymptotic probability calculations, we can assume that

$$(2.1) \qquad X_{n:i} = \theta + \{(Z_n + \dots + Z_i)/\delta cn\}^{\delta},$$

where $Z_1, Z_2, \ldots$ are independent and exponentially distributed random variables with mean 1, $i = 1, 2, \ldots, k(n)$, $\lim_{n\to\infty} k(n) = \infty$, $\lim_{n\to\infty} k(n)/n^\epsilon = 0$ for all $\epsilon > 0$. Let $m(n)$ be an integer, $m(n) < k(n)$ and define

$$\hat{\delta}_{m(n)} = \log \frac{X_{n:k(n)} - X_{n:1}}{X_{n:m(n)} - X_{n:1}} \Big/ \log \frac{k(n)}{m(n)}.$$

Using the representation (2.1) and the law of large numbers, for every sequence $m(n)$ for which

$$\lim_{n\to\infty} m(n) = \infty \text{ and } \lim_{n\to\infty} (k(n) - m(n)) = \infty \tag{2.2}$$

the estimator $\hat{\delta}_{m(n)}$ is consistent for δ. Clearly $\hat{\delta}$ in (1.2) is a special case of $\hat{\delta}_{m(n)}$ with $m(n) = [k(n)/2]$. Although this choice of $m(n)$ is intuitively appealing, it is not a very good one. Suppose that

$$\lim_{n\to\infty} \frac{m(n)}{k(n)} = p \quad (0 < p < 1). \tag{2.3}$$

Then, it is not hard to show that the asymptotic variance of $\hat{\delta}_{m(n)}$ is given by

$$\lim_{n\to\infty} k(n) \operatorname{Var} \hat{\delta}_{m(n)} = \delta^2(1-p)/(p \log^2 p) \equiv \delta^2 V(p). \tag{2.4}$$

The function $V(p)$ has a J shape on (0, 1) and attains its minimum at $p_0 = .2032$. Hence, in the sense of minimum variance, the optimal choice of $m(n)$ will be $m(n) = [p_0 k(n)] \approx k(n)/5$.

Finally, the results hold under milder conditions than considered by Weiss [1]. Namely, it is enough to assume that F is regularly varying at θ:

$$\lim_{y \downarrow 0} \frac{F(\theta + ty)}{F(\theta + y)} = t^\gamma \text{ for all } t > 0.$$

3. CONCLUDING REMARKS

The shape-parameter $\delta = 1/\gamma$ is a scale parameter for $\hat{\delta}_{m(n)}$, whose distribution does not depend on θ or c. Thus, $\hat{\delta}_{[p_0 k(n)]}$ is optimal (among all $\hat{\delta}_{[pk(n)]}$, $0 < p < 1$) uniformly in θ, c and δ. The optimality is preserved if we use $1/\hat{\delta}_{[pk(n)]}$ to estimate γ. The asymptotic relative efficiency (ARE) of $\hat{\delta}_{[pk(n)]}$ with respect to $\hat{\delta}_{[p_0 k(n)]}$ is at least 90% in the interval $p \in [.1, .35]$ and only 74.2% for $p = .5$ (see Table 1).

REFERENCE

[1] Weiss, L., "Asymptotic Inference about a Density Function at and End of Its Range," Naval Research Logistics Quarterly, *18*, 111-114 (1971).

TABLE 1 — *The asymptotic relative efficiency of* $\hat{\delta}_{[pk(n)]}$ *with respect to* $\hat{\delta}_{[p_0k(n)]}$ *(in percents)*

p	ARE	p	ARE
0	0	.20	100.0
.01	33.1	.30	95.9
.02	48.2	.40	86.4
.03	58.7	.50	74.2
.04	66.7	.60	60.4
.05	72.9	.70	45.8
.06	78.0	.80	30.8
.07	82.2	.90	15.4
.08	85.7	.95	7.7
.09	88.5	.99	1.5
.10	91.5	1.00	0

A BRANCH-AND-BOUND ALGORITHM FOR SOLVING FIXED CHARGE PROBLEMS

Patrick G. McKeown

College of Business Administration
The University of Georgia
Athens, Georgia

ABSTRACT

Numerous procedures have been suggested for solving fixed charge problems. Among these are branch-and-bound methods, cutting plane methods, and vertex ranking methods. In all of these previous approaches, the procedure depends heavily on the continuous costs to terminate the search for the optimal solution. In this paper, we present a new branch-and-bound algorithm that calculates bounds separately on the sum of fixed costs and on the continuous objective value. Computational experience is shown for various standard test problems as well as for randomly generated problems. These test results are compared to previous procedures as well as to a mixed integer code. These comparisons appear promising.

1. INTRODUCTION

One class of mathematical programming problems that has been of continuing interest to researchers is the linear fixed charge problem (LFCP). The LFCP is similar in structure to a minimization linear programming problem except that a fixed or lump charge must be paid if the associated continuous variable is positive. It is this discontinuity that has made the LFCP difficult to solve.

The LFCP may be formulated as follows:

$$
\begin{array}{lll}
(P) & \text{maximize} & C^T x + F^T y \\
 & \text{subject to} & x \in S \\
 & \text{and} & y_j = \begin{cases} 1 \text{ if } x_j > 0 \\ 0 \text{ if } x_j = 0 \end{cases} \\
 & \text{where} & S = \{x | Ax = b,\ x \geqslant 0\}.
\end{array}
$$

If n' is the number of structural variables and l is the number of slack and/or surplus variables, then $n = n' + l$ is the number of x variables. Then A is $m \times n$; C is $n \times 1$; F is $n \times 1$ and nonnegative. Both C and F have the last l components equal to zero. Finally, b is a nonnegative $m \times 1$ vector. The c_j's are the continuous or "unit" costs and the f_j's are the fixed costs. Also, let $N = \{1, \ldots, n\}$ and $M = \{1, \ldots, m\}$.

Previous work on this problem has centered about either the LFCP with a general constraint matrix or the fixed charge transportation problem (FCTP) (e.g. [5]). The latter has the special form of the Hitchcock transportation problem. In all cases, work has sought to exploit the result first shown by Hirsch and Dantzig [4] that the optimal solution to any fixed charge problem (if it exists) will occur at an extreme point of the constraint set. It is not as easy as it may seem to find a global minimum since it has also been recognized that many basic solutions are local minima [7].

Work on the general version LFCP has taken three different directions. Steinberg [11] used a search procedure to solve problems with as many as 30 variables and 15 constraints while Taha [13] used an adjacent vertex cutting plane procedure to solve problems of size 15 × 20. Finally, McKeown [8] used a vertex ranking approach first suggested by Murty [9] to solve both general and transportation fixed charge problems. He solved general problems with as many as 20 structural variables.

Our approach to the problem will be to use a branch-and-bound procedure to solve the LFCP. We will discuss this algorithm in Section 2 and present a numerical example in Section 3. Finally, we present and discuss computational results in the last section.

2. THE BRANCH-AND-BOUND ALGORITHM

Our branch-and-bound algorithm is a straightforward application of the procedure described by Geoffrion and Marsten [3]. They suggest that difficult problems such as (P) can be solved by a systematic process made up of separation, relaxation, and fathoming. The separation generates candidate problems (CPs) which are added to a candidate list (CL). A (CP) is removed from the (CL) and an attempt is made to find a bound on its optimal solution. Usually, a relaxation of the (CP) is solved to calculate this bound. If the bound is worse than the value of a feasible solution or if the (CP) is infeasible, it is possible to eliminate the (CP). If it cannot be eliminated, the (CP) is then separated also. If the (CP) is discarded, it is said to be *fathomed*. This procedure continues until the (CL) is empty. At any point in the procedure, the best feasible solution found to date is known as the *incumbent*.

The unique part of our branch-and-bound algorithm is based upon the following propositions:

PROPOSITION 1: Consider Problem (P_c) below (the continuous portion of (P)):

$$(P_c) \quad \text{minimize} \quad C^T x$$
$$\text{subject to} \quad x \in S$$

Then, if (x^*, y^*) solves problem (P) and $\hat{x}$ solves problem (P_c), then $C^T\hat{x} \leqslant C^T x^*$.

PROOF: Trivial from the theory of linear programming.

PROPOSITION 2: Consider problem (P_δ) below

$$(P_\delta) \quad \text{minimize} \quad \sum_{j \in N} f_j y_j$$

$$\text{subject to} \quad \sum_{j \in N} \delta_{ij} y_j \geqslant \Gamma_i \; i \in M,$$

$$y_j \geqslant 0, \; j \in N$$

$$\text{where} \quad \delta_{ij} = \begin{cases} 1 \text{ if } a_{ij} > 0 \\ 0 \text{ if } a_{ij} \leqslant 0 \end{cases} \text{ in } A \text{ and } \Gamma_i = \begin{cases} 1 \text{ if } b_i > 0 \\ 0 \text{ otherwise} \end{cases} \; i \in M.$$

Then if (x^*, y^*) solves problem (P) and $\hat{y}$ solves problem P_δ, then $F^T\hat{y} \leqslant F^Ty^*$.

PROOF: See [8]. These two propositions give rise to the following result which defines the relaxation of problem (P):

PROPOSITION 3: If $v(P_c) = C^T\hat{x}$, $v(P_\delta) = F^T\hat{y}$, and $v(P) = C^Tx^* + F^Ty^*$, then

$$v(P_c) + v(P_\delta) \leqslant v(P).$$

PROOF: Follows from previous results.

Separation occurs by selecting a variable x_j which is basic and nondegenerate in (P_c) and which has not been previously used for separation, and alternately setting the corresponding y_j value to 0 and 1. This avoids problems with infeasibility since we know that any solution to (P_c) is also a solution to (P). We also avoid degeneracy considerations by this selection of variables.

It is not necessary to solve (P_c) and (P_δ) from scratch at each new separation. This may be avoided by using the optimal tableau for each of these problems corresponding to the (CP) being separated. If y_j is set to be 0, this implies that x_j equals 0 also. The new tableaus for this branch for (P_c) and (P_δ) can be found by using the dual simplex algorithm to drive y_j and x_j, respectively, out of the optimal basis for each problem. If, in either case, no feasible solution exists with $x_j = 0$ and $y_j = 0$, then the branch is terminated. Otherwise $v(P_c)$ and $v(P_\delta)$ are equal to the optimal solution values of the respective problems. Obviously, if y_j is not basic in (P_δ) for the (CP), then $v(P_\delta)$ does not change.

On the other hand, if y_j is set to 1 then x_j is already basic and positive by our selection of separation variables so no action is necessary on (P_c). Hence, the optimal tableau for the (CP) is also optimal for the new problem formed by the separation. In the case of (P_δ), we can compute the optimal solution for this branch by noting that (P_δ) is in actuality a relaxation of a set covering problem. We can then determine the effect of setting $y_j = 1$ by seeing that the new set to be covered is equal to $1 - \delta_{ij}$ for $i \in M$. We are then interested in solving the following modified version of (P_δ) which we will term problem $(P1_\delta)$:

$$(P1_\delta) \quad \text{minimize} \quad \sum_{j \in J_2} f_j y_j$$

$$\text{subject to} \quad \sum_{j \in J_2} \delta_{ij} y_j \geqslant \Gamma_i - \sum_{j \in J_1} \delta_{ij} \; i \in M,$$

$$y_j \geqslant 0 \qquad j \in J_2,$$

where $J_0 = \{j \mid y_j = 0\}$

$J_1 = \{j \mid y_j = 1\}$

and $J_2 = N - \{J_1 U J_0\}$.

It is not necessary to solve $(P1_\delta)$ anew at each separation of a (CP). If, for the (CP), $\bar{B}_\delta$ is the optimal basis for $(P1_\delta)$, $\bar{\Gamma}$ is the right-hand-side, and Δ_j is the jth column of the constraint matrix, i.e. $\{\delta_{ij}, i \in M\}$, for the separation variable, then $\bar{B}_\delta^{-1}(\bar{\Gamma} - \Delta_j) = B_\delta^{-1}\bar{\Gamma} - B_\delta^{-1}\Delta_j$ is the new right-hand-side for $(P1_\delta)$ for the $y_j = 1$ branch. This allows us to compute a new right-hand-side for $(P1_\delta)$ at this branch by subtracting the jth transformed column from the right-hand-side corresponding to the (CP) being separated. If the new right-hand-side is feasible ($\geqslant 0$), then the present basis $(\bar{B}_\delta)$ is optimal. If this is not true, then the dual simplex method can be used to attempt to achieve a feasible solution. If no basic feasible solution exists for the new right-hand-side, this branch can be terminated.

If Z_δ is the optimal solution value to $(P1_\delta)$ as found above, then $v(P_\delta) = Z_\delta + \sum_{j \in J_1} f_j$. Also, if $\bar{Z}$ is the value of an incumbent solution and for any branch, $v(P_\delta) + v(P_c) \geqslant \bar{Z}$, the branch can be terminated.

This process begins by solving the original versions of (P_δ) and (P_c) to determine initial optimal tableaus. These tableaus are then used to generate successive tableaus as described earlier. Rather than storing the entire tableau for each (CP) in the (CL), the index sets of the basic variables for (P_c) and (P_δ), which we denote as B_c and B_δ, are stored as binary words. When the (CP) is removed, a "crashing" routine is used to compute the desired tableaus for (P_c) and (P_δ) corresponding to the index sets. *By crashing, we mean forcing the variables in* the index set into the current basis without regard to intermediate feasibility. The fact that the current and desired tableaus are both feasible insures that the desired tableau will be achieved.

A Linear Programming Lower Bound

Problem (P) may also be formulated as a mixed integer programming problem as follows:

$$
\begin{array}{lll}
(P_{MI}) \quad \text{minimize} & Z = C^T x + F^T y & \\
\qquad\quad \text{subject to} & x \in S & \\
& u_j y_j - x_j \geqslant 0 & j \in N \\
& y_j \in \{0, 1\} & j \in N
\end{array}
$$

where $u_j \geqslant x_j$ for $j \in N$.

For fixed charge transportation problems with supplies $\{S_i\}$ and demands $\{D_j\}$, then $\mu_{ij} = \min\{S_i, D_j\}$. Or, if the A matrix is *nonnegative*, the u_j values can be found by

$$u_j = \max_{a_{ij} > 0} \left\{ \frac{b_i}{a_{ij}} \right\}.$$

However, if the A matrix does not satisfy either of the above criteria, then we must compute the u_j values by other means. To do this let $\bar{Z}$ be the the value of an incumbent solution. Then we can require that $v(P_c) + v(P_\delta) \leqslant \bar{Z}$. We can use this result to compute upper bounds on the x_j using the following linear programming problem:

$$\begin{array}{lll} (P_u) & \text{maximize} & x_j \\ & \text{subject to} & x \in S \\ & & C^T x \leqslant \bar{Z} - v(P_\delta). \end{array}$$

For each x_j, $u_j = v(P_u)$.

If we relax the integer restriction on the y_j variables to $0 \leqslant y_j \leqslant 1$ in P_{MI}, we have a linear programming relaxation which yields an optimal solution value Z_l. Obviously, $Z_l < \bar{Z}$, so we could use this as a lower bound in a branch-and-bound procedure. This bounding procedure will be compared in a later section to $v(P_c) + v(P_\delta)$.

3. EXAMPLE

As an illustration of our branch-and-bound solution procedure for the linear fixed charge problem, consider the following problem:

$$\begin{array}{lll} \text{minimum} & 4x_1 + 2x_2 + 3x_3 + 24y_1 + 12y_2 + 16y_3 & \\ \text{subject to} & x_1 + 3x_2 & \geqslant 15 \\ & x_1 \qquad\quad + 2x_3 & \geqslant 10 \\ & 2x_1 + x_2 & \geqslant 20 \\ & y_j = \begin{cases} 1 \text{ if } x_j > 0 & j = 1, 2, 3 \\ 0 \text{ if } x_j = 0 & \end{cases} & \\ & x_j \geqslant 0, \quad j = 1, \dots, 3. & \end{array}$$

For this problem P_δ is as follows:

$$\begin{array}{lll} \text{minimum} & 24y_1 + 12y_2 + 16y_3 & \\ \text{subject to} & y_1 + y_2 & \geqslant 1 \\ & y_1 \qquad\quad + y_3 & \geqslant 1 \\ & y_1 + y_2 & \geqslant 1 \\ & y_j \geqslant 0, \quad j = 1, 2, 3. & \end{array}$$

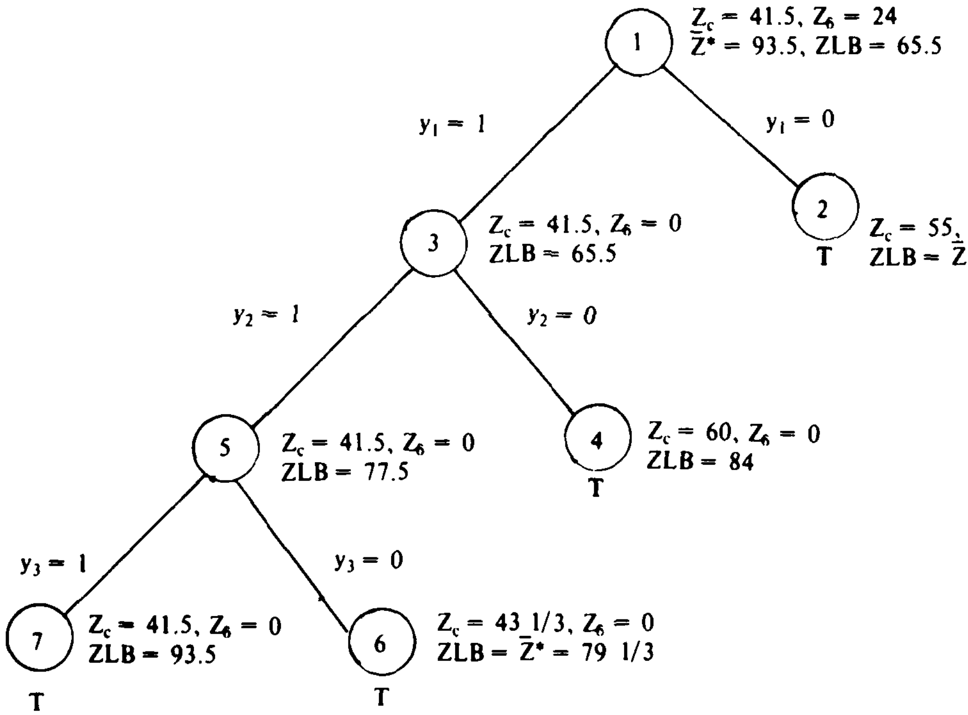
Solving P_c and P_δ we get basic variable index sets $B_c = \{1, 2, 3\}$ and $B_\delta = \{1\}$, $v(P_c) = Z_c = 41.5$, $v(P_\delta) = Z_\delta = 24$. We can also compute a feasible solution to P using B_c to obtain a $\bar{Z}^* = 93.5$. We now branch on this solution to P_c by setting $y_1 = 0$ and $x_1 = 0$. This yields a $Z_c = 55$ and $Z_\delta = 28$ for a lower bound of 83 which in this case also equals a new $\bar{Z}^*$ value at this solution (node 2 in diagram).

We now set $y_1 = 1$ and compute $Z_c = 41.5$, $Z_\delta = 0$, $ZLB = 41.5 + 0 + 24 = 65.5$, $B_c = \{1, 2, 3\}$, $B_\delta = \{\emptyset\}$, $J_0 = \emptyset$, $J_1 = \{1\}$, and $J_2 = \{2, 3\}$. We now branch on this last node by setting $y_2 = 0$ and $y_2 = 1$. The zero node is terminated since $Z_c = 60$, $Z_\delta = 0$ and

$ZLB = 60 + 24 = 84 \geqslant 83$. The one node yields a $Z_c = 41.5$, $Z_\delta = 0$, and $ZLB = 41.5 + 24 + 12 = 77.5$. For this last node (5 on the diagram), $B_c = \{1, 2, 3\}$, $B_\delta = \{4, 6\}$, $J_0 = \emptyset$, $J_1 = \{1, 2\}$, and $J_2 = \{3\}$.

If we then branch on node 5 by setting $Y_3 = 1$, we determine a $ZLB = 93.5$ which is greater than 83 so we terminate this branch. However, branching on $y_3 = 0$ yields a new $\bar{Z}^* = 79\ 1/3$. $Z_c = 43\ 1/3$, $Z_\delta = 0$ and $ZLB = 43\ 1/3 + 24 + 12 = 79\ 1/3$. Since the lower bound equals $\bar{Z}^*$, we may also terminate this node with the present incumbent solution, i.e., $\bar{Z}^* = 79\ 1/3$ and $B_c = \{1, 2, 6\}$, being optimal.

At the root node, the optimal solution to the LP relaxation of (P_{MI}) is found by computing $u_j = 15, 20, 5$ for $j = 1, 2, 3$. The value of Z_l using these upper bounds for this problem is 58.7 as compared to the lower bound computed above of $65.5 = v(P_c) + v(P_\delta)$.



4. COMPUTATIONAL RESULTS

The branch-and-bound algorithm discussed previously was programmed in FORTRAN for testing on a CDC CYBER 70/74 computer. In coding the algorithm, standard FORTRAN was used with the exception of bit manipulation extensions, i.e., storage of all pertinent variable lists was done using bit manipulation of the CYBER 60 bit words.

To determine a good starting candidate for optimality, Phase I of Walker's [15] heuristic was used after the solution of the continuous portion of the linear fixed charge problem. This has been shown to give very good approximate results. The procedure for selecting a node to remove from the partial solution list to branch on was very simple in that we always chose the last node on the list. Similarly, we used a simple approach to select a variable to use in branching. The branching procedure always chose the first free variable in the basis for the continuous problem to branch upon by first setting the corresponding integer variable to zero and then setting it to one. From this discussion, it is clear we have not attempted to "optimize" our algorithm.

To test this code, various types of test problems were solved. The first class of problems were general fixed charge problems with all equality constraints. These problems were originally generated by Cooper and Drebes [1] and have been used by various authors as benchmark problems. These test problems are of size 5 × 10 with the following characteristics:

(i) $|a_{ij}| \leqslant 20$

(ii) $1 \leqslant b_i \leqslant 999$

(iii) $1 \leqslant c_j \leqslant 20$

(iv) $1 \leqslant f_j \leqslant 999$

and are 50% dense. To test larger problems, these 5 × 10 problems were combined by placing them on the diagonal to generate 10 × 20 and 15 × 30 problems. The results of solving these problems with our code are shown in Table 1.

TABLE 1 — *Computational Results With Standard Test Problems*

Algorithm	Machine	$m \times n$	Number of Problems	Average Solution Time
Branch-and-Bound	CYBER 70/74	5 × 10	12	.114
		10 × 20	6	1.866
		15 × 30	4	15.189
Steinberg's Search	IBM 7072	5 × 10	15	9.600
		15 × 30	10	1266.0
Vertex Ranking	UNIVAC 1108	5 × 10	9	.496
		10 × 20	5	37.704
		15 × 30	1	*
Walker	CDC 1604	5 × 10	52 (52)	4.
		15 × 30	5 (5)	18.

*Storage overflow

As noted above, these problems have been used as test problems by various previous researchers. For that reason, we have attempted to supply comparative computational results. Previous results shown in Table 1 are those of Steinberg's testing of his search procedure on the IBM 7072 and McKeown's testing of the vertex ranking procedure on the UNIVAC 1108. In each case we have shown the CPU seconds to find and prove an optimal solution. In the case of our branch-and-bound code, these times are based on using the FTN compiler with $OPT = 2$ on a CYBER 70/74. We have also shown results published by Walker for his complete heuristic procedure. For that algorithm, the value in parentheses refers to the number of problems tested that were optimal.

In addition to the benchmark problems we have compared our branch-and-bound code for fixed charge problems (FCBB) to a state-of-the-art mixed-integer code based on Beale and Tomlin [14] work as implemented by Sinha [12] which we will denote as SMIP. To use SMIP, upper bounds on each variable must be found. These were found by solving P_u for each variable. This was done in FCBB using the heuristically derived value of $\bar{Z}^*$ and the linear programming solution value of P_8, i.e., $v(P_8)$, at the root node. Since the constraint set remains the same for each variable, it was not necessary to solve each version of P_u from scratch. Instead, the optimal basis for the first variable problem was used as a starting point for the second variable problem, and so on until all upper bounds were computed.

The problems solved were the same as the Cooper-Drebes problems with the exception that problems of various sizes were generated. The results of this testing are shown in Table 2. For each set of five problems, we have shown the average time and range of times (CPU seconds) for FCBB and SMIP. We have also shown these same values for the FCBB without the Walker heuristic and for SMIP without including the times necessary to calculate the upper bounds. The former calculations were made to determine the effect on FCBB of not using a heuristic since none was used in SMIP. Similarly, the latter calculations were made to compare FCBB and SMIP under the assumption that upper bounds were known for each variable and did not need to be calculated.

Looking at Table 2, we note that for Problem Sets 1-5, FCBB is significantly faster than SMIP for 5 × 10 problems and this disparity increases as the number of variables increases until FCBB is faster by more than a factor of 20 for 5 × 50 problems. The same general trend exists when FCBB is compared to SMIP without the upper bound calculations but as would be expected the SMIP times are slightly faster. For problem sets 6-9, the relative efficiency of FCBB compared to SMIP is not as dramatic but FCBB remains at least twice as fast as SMIP. Also, SMIP exceeded storage limitations for the largest problems.

Two other results may also be noted from Table 2. First, for problem sets 1-5, increasing the number of variables has some effect on solution times for FCBB but not as much as would be expected. However, there is a much more dramatic effect of increasing the number of variables for problem sets 6-9. The second result is that removing the Walker Heuristic has very little effect on the solution times except for the 5 × 30 problems. This indicates that the algorithm is finding a good incumbent value early in the branching process.

One of the reasons for the relative efficiency of FCBB to SMIP is that while the LP relaxation has $m + n$ constraints. Since both procedures use pivoting to calculate bounds, the larger number of constraints slows the LP relaxation.

TABLE 2 — *Comparison of Algorithms (all ≥ constraints)*

Problem Set	Size	FCBB Time		SMIP Time		FCBB (without heuristic) Time		SMIP (minus Bound Calculation Time)	
		Ave.	Range	Ave.	Range	Ave.	Range	Ave.	Range
1	5 × 10	.070	.023- .219	.202	.130 .389	.068	.022- .531	.136	.086- 2.482
2	5 × 20	.151	.043- .500	1.018	.444- 2.805	.158	.045- .531	.815	.296- 2.482
3	5 × 30	.222	.087 .610	1.896	.785- 3.255	.320	.103- 1.048	1.523	.613- 2.813
4	5 × 40	.349	.136- .549	5.028	2.593- 8.678	.394	.239- .582	4.375	2.036- 8.038
5	5 × 50	.323	.172- .668	7.055	4.492- 13.383	.362	.166- .661	6.088	3.656- 12.388
6	10 × 20	.424	.365- .626	1.993	1.082- 3.087	.487	.210- .817	1.295	.541- 2.135
7	10 × 30	2.554	1.086- 4.373	6.416	3.499- 10.818	—	—	4.697	2.219- 9.041
8	10 × 40	7.273	2.111- 20.001	15.727	10.079 25.011	—	—	12.767	7.383- 21.632
9	10 × 45	11.822	1.793- 33.986	+	+	—	—	+	+

+ Exceeded storage limitations.

The final result of interest in Table 2 is the wide range of times for each problem set. This is similar to results reported by Ross and Soland [10] for Special Transportation Problems and by Fisk and McKeown [2] for Pure Fixed Charge Transportation Problems. It is unclear in any of these contexts as to what makes a problem "easy" or "difficult" to solve.

Table 3 gives results of using FCBB to solve six versions of problem set 3. Three characteristics were varied to determine their effect on solution times. These characteristics were density, constraint type, and relative size of fixed and continuous costs. With the exception of these three characteristics, the problems in Table 3 are the same as Problem Set 3 in Table 2, i.e., five problems in each set with each problem being 5 × 30 and being randomly generated using the same parameters as the Cooper-Dreber problems. In each case, the original results for Problem Set 3 are used as a bench mark. To summarize, Table 3 shows that problems with equality constraints are more difficult to solve than problems with greater-than-or-equality constraints but problems that are 90% dense are easier to solve than problems that are 50% dense. Finally, the greater the continuous costs relative to the fixed costs, the easier the problems are to solve.

The first two results have not been reported before while the last is similar to that shown by Kennington [6] for Fixed Charge Transportation Problems. The first result is thought to be caused by the low value of P_8 as compared to the sum of fixed charges in the optimal solution

TABLE 3 — *Effect of Parameter Variation*

Problem Set	Size	Density	C_j Range	F_j Range	Constraint Type	FCBB[a] Ave. Time	Time[a] Range
10	5 × 30	50%	1-20	0-999	GT	100	39.18-274.77
11	5 × 30	50%	1-20	0-999	EQ	402.25	100.00-819.36
12	5 × 30	90%	1-20	0-999	GT	78.82	64.86-91.89
13	5 × 30	50%	0-0	0-999	GT	276.57	80.63-561.26
14	5 × 30	50%	1-99	0-999	GT	79.81	38.28-102.70
15	5 × 30	50%	1-20	0-20	GT	38.28	32.88-47.74

[a]All times as a percent of Average Time for Problem Set 10 [Same as Set 3].

to P. On the other hand, higher density values will lead to higher values of P_δ relative to the sum of fixed charges in the optimal solution. Finally, higher continuous costs lead to a better overall bound since $v(P_c)$ tends to dominate the overall cost structure and the solution to P_c comes closer to matching the optimal solution to P. However, FCBB was able to solve problems with the continuous costs equal to zero in a reasonable length of time. This is the first time problems of this type have been attempted and successfully solved. For problems of this type, the P_δ bound alone appears to be adequate.

In Table 4, we have compared the lower bounds found by FCBB and by using the optimal solution to the linear programming relaxation (P_{MI}). We used six problems previously solved as portions of Problem Sets 3 and 8. For each problem, we have shown the time to compute the bound, the lower bounds as a percentage of the optimal solution ($V(P)$) and the solution time. The results from these six problems do not show any clear trends. The lower bounds tend to favor the FCBB algorithm as do the solution times but the lower bound for Problem 3c is better using the LP relaxation. The only general result that can be derived from the table is that the quality of the bound does not in itself guarantee a faster solution time for a given problem.

TABLE 4 — *Comparison of Bounds*

Problem	Size	FCBB[a] Bound Time	FCBB[b] Bound	FCBB[c] Solution Time	LP[a] Bound Time	LP[b] Bound	LP[c] Solution Time
3a	5 × 30	38.59	77.98	5.64	39.86	72.65	100
3b	5 × 30	63.95	85.31	11.05	79.30	71.99	100
3c	5 × 30	58.09	71.01	11.69	81.41	77.09	100
8a	10 × 30	12.82	54.29	14.28	23.64	29.26	100
8b	10 × 30	9.92	60.85	40.95	49.95	53.84	100
8c	10 × 30	5.25	35.47	52.81	41.59	45.21	100

[a]All times as a percent of time to find the optimal solution for that algorithm. For the LP bound, the times include the time to compute upper bounds on the variables.

[b]All bounds as a percent of optimal value ($v(P)$).

[c]All solution times as a percent of time using LP relaxation (SMIP) to find optimal solution.

5. CONCLUSIONS

We have presented here a branch-and-bound procedure for solving linear fixed charge problems that uses a new approach to bounding on the fixed charges as well as on the continuous costs while not requiring a knowledge of upper bounds on the continuous variables. Computational results for this procedure have been presented which appear to be significantly better than previous algorithms or state-of-the-art mixed integer codes.

Future research in this area would include testing of various selection rules and branching schemes in order to optimize the algorithm, use of mass storage files to store bases for partial solutions and testing to determine what makes a problem "easy" or "difficult" to solve. This procedure could also be specialized for fixed charge transportation problems by using network dual pivoting procedures for calculating bounds.

REFERENCES

[1] Cooper, L. and C. Drebes, "An Approximate Solution Method for the Fixed Charge Problem," Naval Research Logistics Quarterly, *8*, 101-113 (1967).

[2] Fisk, J. and P. McKeown, "The Pure Fixed Charge Transportation Problem," Naval Research Logistics Quarterly, *26*, 631-641 (1979).

[3] Geoffrion, A.M. and R.E. Marsten, "Integer Programming Algorithms: A Framework and State-of-the-Art Survey," Management Science, *18*, 465-491 (1972).

[4] Hirsch, W.M. and G.B. Dantzig, "The Fixed Charge Problem," Naval Research Logistics Quarterly, *15*, 413-424 (1968).

[5] Kennington, J.L. and V.E. Unger, "A New Branch-and-Bound Algorithm for the Fixed Charge Transportation Problem," Management Science, *22*, 1116-1126 (1976).

[6] Kennington, J.L., "The Fixed-Charge Transportation Problem: A Computational Study with a Branch-and-Bound Code," AIIE Transactions, *8*, 241-247 (1976).

[7] McKeown, P. "An Extreme Point Ranking Algorithm for Solving the Linear Fixed Charge Problem," Ph.D. Dissertation, University of North Carolina, Chapel Hill, NC (1973).

[8] McKeown, P. "A Vertex Ranking Procedure for Linear Fixed Charge Problems," Operations Research, *23*, 1183-1191 (1975).

[9] Murty, G., "Solving the Fixed Charge Problem by Ranking the Extreme Points," Operations Research, *16*, 268-279 (1968).

[10] Ross G.T. and R. Soland, "A Branch and Bound Algorithm for the Generalized Assignment Problem," Mathematical Programming, *8*, 91-103 (1975).

[11] Steinberg, D.I., "The Fixed Charge Problem," Naval Research Logistics Quarterly, *17*, 217-236 (1970).

[12] Sinha, P., Private Communication (1977).

[13] Taha, H., "Concave Minimization Over a Convex Polyhedron," Naval Research Logistics Quarterly, *20*, 533-547 (1973).

[14] Tomlin, J.A., "Branch and Bound Methods for Integer and Nonconvex Programming" in: Integer and Nonlinear Programming, 437-450, J. Abadie, Editor, American Elsevier Publishing Company, New York (1970).

[15] Walker, E., "A Heuristic Adjacent Extreme Point Algorithm for the Fixed Charge Problem," Management Science, *22*, 587-596 (1976).

FLUCTUATIONS OF THE OPTIMAL ADVERTISING AND INVENTORY POLICY: A QUALITATIVE ANALYSIS

Yves Balcer

Department of Economics
University of Wisconsin
Madison, Wisconsin

ABSTRACT

This paper discusses the properties of the inventory and advertising policy minimizing the expected discounted cost over a finite horizon in a dynamic nonstationary inventory model with random demand which is influenced by the level of promotion or goodwill. Attention is focused on the relation between the fluctuations over time of the optimal policies and the variations over time of the factors involved, i.e., demand distributions and various costs. The optimal policies are proved to be monotone in the various factors. Also, three types of fluctuations over time of the optimal policies are discussed according to which factor varies over time. For example, if over a finite interval, the random demand increases (stochastically) from one period to the next, reaches a maximum and then decreases, then the optimal inventory level will do the same. Also the period of maximum of demand never precedes that of maximum inventory. The optimal advertising behaves in the opposite way and its minimum will occur at the same time as the maximum *of the inventory. The* model has a linear inventory ordering cost and instantaneous delivery of stocks; many results, however, are extended to models with a convex ordering cost or a delivery time lag.

1. INTRODUCTION

This paper focuses on the problem faced by the manager of a store with important inventory costs and an advertising budget in the presence of fluctuations in economic conditions. In particular, qualitative results are provided linking fluctuations in economic conditions (demand, holding costs, advertising costs) to fluctuations in jointly managed optimal inventory orderings and advertising levels. Under the usual convexity assumptions, knowledge of the fluctuations of the demand over time is sufficient to determine those of the optimal policy without computing them explicitly.

Very few authors have attempted to optimally integrate inventory policy with advertising policy. In fact, the only reference we know of is Miercourt [8]. He established the existence and the characterization of the optimal policy, though he did not attempt to discuss their fluctuations over time. The general approach to the fluctuation problem is along the lines of Karlin's work [5] and its extensions by Veinott [19], [22]. Karlin established the relationship over time of the optimal inventory policy with the fluctuations of the demand for a discrete time dynamic inventory model involving a single commodity with random demand. Veinott extended these results in various ways emphasizing refinements that are possible with translations of the demand distributions. He also showed that the proper concept linking the variations of the optimal policy over time with the fluctuations of the demand and other parameters is the myopic policy. A proper interpretation of some of our results extends the work of Pierskalla [10] relating optimal inventory policy with monotone obsolescence probabilities. The work of Topkis and Veinott [18], Topkis [17] and Veinott [26] on subadditive functions on sublattices and its application to inventory problems by Veinott [25], provides us with the proper tools and methods for our analysis. A knowledge of their theory, at least to the extent of the Appendix of Balcer [3], is essential to understand this paper. Results from this Appendix are referred to by the letter A followed by a number, such as Theorem A2, Lemma A1 Example A5. Results established in this paper are referred to by numbers, such as Theorem 6 or Lemma 5.

An abundant literature (see Balcer [2]) covers the problem of optimizing advertisement expenditures given a known demand-advertising relationship and no inventories. Most of them generalize the model of Arrow and Nerlove [1], which introduces the concept of goodwill increased by advertising and decaying exponentially. In that literature, the work of Kuehn [7] stands out because of the resemblance of some of his results with ours. His model is not clearly defined and is said to be inspired by a previous model of his [6] which was dynamic with random demand (the randomness was introduced by a Markov process via brand switching). He concluded that the advertising expenditure fluctuates over time in unison with sales and with peaks and bottoms of advertising expenditure preceding those of sales.

As mentioned earlier, our main focus is to link these fluctuations with variations in primary factors such as demand, ordering cost, holding and shortage costs, promotion effect, etc. In Section 2, we set up an inventory model with advertisement. For this model some relevant results from Balcer [3] regarding the existence and the characterization of the optimal policy are reproduced.

In Sections 3 and 4 we establish that the optimal solutions in a given period are monotone in the various costs of that period or future ones. In Section 5 we exhibit the relationship between optimal inventory policies which take into account all present and future costs, and myopic inventory policies which take into account only present costs. In Sections 6 and 7 the fluctuations in the optimal inventory and advertising policies are linked to the fluctuations in the myopic policies which, in turn, vary with changes in the parameters of the problem such as the various costs and the demand. The problem of inventory with varying obsolescence probabilities treated previously by Pierskalla is discussed. Finally, in Section 8, the results of the previous sections are extended to models with convex inventory ordering costs or with lag in delivery of the ordered inventory. Sections 5, 6, and 7 are more closely related to Karlin's work and Sections 3, 4, 5 and 8 to Veinott's.

2. MODEL

In this paper we will study a discrete time dynamic model of single commodity management, when the nonnegative demand for the commodity in each period is uncertain but has a known distribution depending on the existing level of goodwill. At the beginning of each period, the manager knows the initial inventory x, the present and future demand distributions and the cost structure. He decides to increase instantaneously the initial inventory x to a starting level $y \geqslant x$ by ordering at a unit cost c, and goodwill to a starting level $b \geqslant 0$ by advertising at unit cost p. The demand $g(b) + U$ is the sum of a nonnegative random variable U and a nondecreasing function of goodwill such that $g(0) = 0$. During this period, the demand $g(b) + U$ occurs, so the terminal inventory becomes $y - g(b) - U$. The initial inventory in the next period is $\eta(y - g(b) - U)$ where $\eta(z) = n_+ z^+ - \eta_- z^-$ with $\eta_+ \geqslant \eta_- \geqslant 0$, $z^+ = \max(z, 0)$ and $z^- = \max(-z, 0)$. When the slope of η is between zero and one, $(1 - \eta_+)$ is interpreted as a depletion or loss of inventory and $(1 - \eta_-)$ as a loss of sales arising from impatient consumers who depart before receiving their orders. The initial level of advertising in the next period is 0 since the effect of advertising is ephemeral.

During each period, the manager incurs a capacity cost, h, on the starting inventory, and a shortage-holding cost, s, on the terminal inventory. Both of these cost functions are convex, and the function s increases to infinity with its argument. Moreover, s is nondecreasing in z on the nonnegative real line. *The function h is bounded below.* There is a unit selling price r. The current sale price is paid when each consumer demand is incurred. This yields a gross revenue of $r(g(b) + U)$ to the manager. If the inventory is positive, the consumer receives the commodity without delay until either the demand is totally satisfied or the inventory is completely exhausted, whichever comes first. If consumers subsequently depart or increase their orders without being served, the manager refunds or pockets, respectively, the then current sale price. The sale is final only when the consumers receive the commodity they have purchased. We assume that EU and $Es(y - g(b) - U)$ are finite, where E denotes the expectation and that the demand function $g(b)$ is increasing and concave in the goodwill. This last assumption has been verified empirically by Shryer [11] and Stone [14] for mail ordering, Telser [15] for cigarettes, Palda [9] for drugs, Clement et al. [4] for milk and Simon [12], [13] for liquor.

Given a finite horizon N, let $\tilde{C}^n(x)$ be the minimum expected discounted cost in the periods $n, \ldots, N$, where x and 0 are respectively the initial inventory and goodwill in period n. The function $\tilde{C}^n$ *can be calculated for each period by the dynamic programming recursion*

$$\begin{aligned}(1) \qquad \tilde{C}^n(x) = \min_{y \geqslant x, b \geqslant 0} \{ & c(y - x) + pb + h(y) + Es(y - g(b) - U) \\ & - r(g(b) + U) + (1 - \eta_-) rE(y - g(b) - U)^- \\ & + \lambda E\tilde{C}^{n+1} [\eta(y - g(b) - U)]\},\end{aligned}$$

for $n = 1, \ldots, N$, where $\tilde{C}^{N+1}(x) = -cx$. Every symbol in (1) should be indexed by n; however, when no confusion is possible because of the context, the index n is suppressed throughout this paper.

It is convenient first to transform the problem to one with no ordering costs, a technique which has been used by Veinott [20]. This is achieved by letting $C^n(x) = \tilde{C}^n(x) + cx$. Second, the problem is not convex in (y, b), but it is in (y, β) where $\beta = g(b)$, the impact of advertising on sales. Because $g(b)$ is increasing in b, the optimal policy in the (y, β)-space can be expressed in the original (y, b)-space. After the two transformations, recursion (1) becomes

$$(2) \qquad C^n(x) = \min_{y \geqslant x, \beta \geqslant 0} \{H(y) + P(\beta) + S(y - \beta) + \lambda EC^{n+1}[\eta(y - \beta - U)]\},$$

where

$$(3) \qquad H(z) = cz + h(z),$$

$$(4) \qquad P(z) = pg^{-1}(z) - rz$$

and

$$S(z) = E[s(z - U) - \lambda c'\eta(z - U) + (1 - \eta_-)r(z - U)^-]. \tag{5}$$

The functions H, P, and S are respectively the aggregate inventory ordering-capacity cost, the aggregate advertising-revenue cost and the aggregate shortage-holding cost. The prime indicates that the parameter is from the following period. Also, a constant term, $-rEU$, has been omitted as it clearly does not affect the choice of the optimal policy. Note now that $C^{N+1}(x) \equiv 0$. For a detailed discussion of these transformations, see Balcer [2] or [3].

In Balcer [3], under the mild condition that S is convex at 0,

$$D^+s(0) + r - \lambda c'\eta_+ \geqslant \eta_-(r - \lambda c') + D^-s(0), \tag{I}$$

where $D^+s(0)$ ($D^-s(0)$) indicates the right (left) derivative of s at zero and some weak conditions at infinity,

$$D^-H(\infty) + D^-S(\infty) > 0, \tag{II.a}$$

$$D^-H(\infty) + D^-P(\infty) > 0, \tag{II.b}$$

$$D^-P(\infty) - D^+S(-\infty) > 0, \tag{II.c}$$

$$D^+H(-\infty) + D^+S(-\infty) < 0, \tag{II.d}$$

it is shown that $C^n(x)$ is convex and nondecreasing in x and that a finite optimal policy exists. Moreover, it is shown that the minimand in (2) is subadditive in (y, β) and that the set of all optimal policies for (2) is a nonempty compact sublattice. For sake of definiteness, the optimal policy $(\bar{y}(x), \bar{\beta}(x))$ is chosen to be the lexicographically least element of that sublattice.

From here on $(\bar{y}^n(x), \bar{\beta}^n(x))$ denotes the optimal policy in period n whose components are the optimal inventory and the optimal controlled demand and which minimizes the right-hand side of (2). The smallest solution to the minimization of the total costs as given by (2) when only one of the two variables can be chosen and when that variable is unconstrained is denoted by $\tilde{y}^n(\alpha)$ and $\tilde{\beta}^n(x)$, respectively. (α is an arbitrary value taken by β when the minimization is done over y only.) Also, the solution to the minimization of the total costs in period n as described by (2) when the two variables are unconstrained is (y^{*n}, β^{*n}). This defines the base stock level in period n (as we shall see below) whose components are the base inventory level and the base controlled demand. The differences, $\bar{y}^n(x) - \bar{\beta}^n(x)$ and $y^{*n} - \beta^{*n}$, are the optimal net inventory and the base net inventory in period n, rewritten $\bar{z}^n(x)$ and z^{*n}, respectively. Finally, the solution to the minimization of the total present costs in period n as given by the right hand side of (2) with the last term omitted is the myopic policy. In the preceeding sentences, if we replace superscript n by subscript n we have the myopic counterparts of the optimal solutions. For example, z_n^* is the myopic base net inventory in period n. In the sequel, as mentioned earlier, the superscript n is omitted when no confusion is possible because of the context.

Finally, the following theorem, which parallels similar results of Veinott [21] for a multiproduct inventory model, holds:

THEOREM 1: Under the above conditions, $\bar{y}(x) = x \vee \tilde{y}(\beta^*)$ and $\bar{\beta}(x) = \tilde{\beta}(y^* \vee x)$ where $\tilde{y}$ and $\tilde{\beta}$ are nondecreasing on their respective domain.

The symbol $\vee$ denotes the maximum of the two elements on each side of it. In addition, if this additional condition holds,

$$\text{(III)} \qquad D^- H(0) + D^- S(0) < 0,$$

the optimal policy is nonnegative and η does not need to be convex.

3. MONOTONICITY OF THE OPTIMAL SOLUTION IN THE AGGREGATE COSTS

In this section, we will describe the changes in $\bar{y}^n(x)$, $\bar{\beta}^n(x)$ and $\bar{z}^n(x)$ when changes occur in the aggregate costs H, P and S. The importance of this and the following section results from the ability to predict changes in the optimal strategy when some underlying conditions are changing without having to recompute the optimal policy. As one is unfamiliar with these aggregate costs, the results of this section will not be as intuitive as those of the next section which are in terms of the primary costs and the demand distributions. We prove all the monotonicity results with the aggregate costs as the proofs are sharper and the results specialize easily to the primary costs.

Before proceeding, additional notations and a lemma on ordered sets of functions—to be defined below—have to be set. The class Ω of all real-valued functions on R can be quasi-ordered by the incremental order defined as follows. First, write $\omega \sim \omega'$ on R if $\omega - \omega'$ is constant on R. Evidently $\sim$ is an equivalence relation on Ω. Now, ω is incrementally smaller than $\omega' \in \Omega$ written $\omega \leqslant \omega'$, if for all $v < w \in R$, $\omega(w) - \omega(v) \leqslant \omega'(w) - \omega'(v)$, if ω and ω' are differentiable, it implies that $\partial\omega(v)/\partial v \leqslant \partial\omega'(v)/\partial v$ for all $v \in R$.

LEMMA 1: If $\emptyset \neq \Omega' \subset \Omega$ and Ω' is a chain, then $\omega(v)$ is superadditive in (ω, v) on $\Omega' \times R$.

PROOF: Suppose (ω, v), $(\omega', v') \in \Omega' \times R$ are incomparable. Since Ω' and R are chains, we may suppose $\omega \leqslant \omega'$ and $v \geqslant v'$. Then $(\omega \wedge \omega')(v \wedge v') + (\omega \vee \omega')(v \vee v') = \omega(v') + \omega'(v) \geqslant \omega'(v') + \omega(v)$ by definition of $\leqslant$, completing the proof.

If ω^n and ω'^n are sequences of real-valued functions, i.e., $\omega^n = (\omega_n, \ldots, \omega_N)$ and $\omega'^n = \omega'_n, \ldots, \omega'_N)$, write $\omega^n \leqslant \omega'^n$ if $\omega_i \leqslant \omega'_i$ for $i = n, \ldots, N$. If Ω is a quasi-ordered set, its dual, Ω^*, contains the same elements with the ordering reversed. In other words, if ω and ω' belong to Ω and $\omega \leqslant \omega'$, then ω and ω' belong to Ω^* and $\omega' \leqslant \omega$.

Denote $\mathscr{H}_n$, $\mathscr{P}_n$ and $\mathscr{S}_n$, the classes of all real valued continuous convex functions H_n, P_n and S_n, respectively, satisfying condition (II). Denote by $\mathscr{H}'_n$, $\mathscr{P}'_n$ and $\mathscr{S}'_n$, respectively, given chains in $\mathscr{H}_n$, $\mathscr{P}_n$ and $\mathscr{S}_n$. The sets $\mathscr{H}^n$, $\mathscr{P}^n$, $\mathscr{S}^n$, $\mathscr{H}'^n$, $\mathscr{P}'^n$, and $\mathscr{S}'^n$ are defined similarly. In the sequel, it is understood that these sets are ordered by the incremental order. In order to exhibit the dependence of $\bar{y}^n(x)$ and $\bar{y}_n(x)$, say, on the aggregate costs, we write $\bar{y}(x; \omega^n)$ and $\bar{y}(x; \omega_n)$ instead of $\bar{y}^n(x)$ and $\bar{y}_n(x)$, respectively, where ω can be either H, P or S. In Theorems 2, 3 and 4 to follow, all parameters of the problem are held fixed but for the one so specified in the theorem. Changes in the optimal policies are linked to the variations of the aggregate ordering-capacity cost H, the aggregate advertising-revenue cost P, and the aggregate shortage-holding cost S. The next four theorems, particularly Theorem 5, are an extension of a result and a proof of Veinott [23] for the impact of future demands on current optimal inventory level; this result was subsequently extended by Topkis [16] for the impact of future cost functions characterized by a single parameter on current optimal inventory level.

THEOREM 2: For all n, $\bar{y}(x;H^n)$ and $\bar{z}(x;H^n)$ are nonincreasing in H^n on $\mathscr{H}^n$, $\bar{\beta}(x;H^n)$ is nonincreasing in H_n and nondecreasing in H^{n+1} on $\mathscr{H}^n$, $\tilde{\beta}(x;H^n)$ is independent of H_n and nondecreasing in H^{n+1} on $\mathscr{H}^n$, and $C(x;H^n)$ is superadditive in (x, H_i) on $R \times \mathscr{H}_i'$ for $i = n, \ldots, N$.

PROOF: See Appendix.

THEOREM 3: For all n, $\bar{y}(x:P^n)$ is nonincreasing in P^n on $\mathscr{P}^n$, $\tilde{\beta}(x;P^n)$ and $\bar{\beta}(x;P^n)$ are nonincreasing in P_n and nondecreasing in P^{n+1} on $\mathscr{P}^n$; $\bar{z}(x;P^n)$ is nondecreasing in P_n and nonincreasing in P^{n+1} on $\mathscr{P}^n$, and $C(x;P^n)$ is superadditive in (x, P_i) on $R \times \mathscr{P}_i'$ for $i = n, \ldots, N$.

PROOF: See Appendix.

THEOREM 4: For all n, $\bar{y}(x;S^n)$ and $\bar{z}(x;S^n)$ are nonincreasing in S^n on $\mathscr{S}^n$, $\bar{\beta}(x;S^n)$ and $\tilde{\beta}(x;S^n)$ are nondecreasing in S^n on $\mathscr{S}^n$; also $C(x;S^n)$ is superadditive in (x, S_i) on $R \times \mathscr{S}_i'$ for $i = n, \ldots, N$.

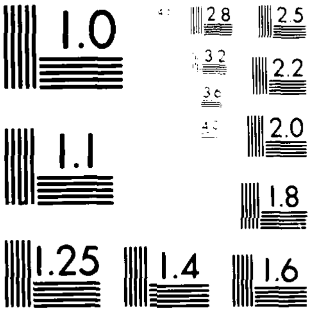
PROOF: See Appendix.

In Theorems 2, 3 and 4 we establish that an increase in any present and future aggregate costs lead to a lower optimal level of inventory on hand both before and after adverstising. This is to be expected, as the cost of capacity goes up, less inventory is purchased in anticipation of sales; also, as the cost of advertising goes up, less advertising will take place, thus less inventory is needed; and finally, as the holding cost function increases, less inventory is pruchased, so less will remain at the end of the period, thus reducing the holding costs. The situation is a bit more delicate for the optimal level of advertising. Recall that $\bar{b}(x) = g^{-1}(\bar{\beta}(x))$, where g^{-1} is monotone increasing. When capacity cost increases in the present period, advertising is reduced as the cost increase has led to a reduced inventory on hand to be sold; but when capacity costs increase for the future periods, advertising is expanded in the current period to increase current demand and therefore reduce the initial inventory level in future period. Obviously, as advertising cost increases in the current period, the amount of advertising is reduced; however, as advertising costs increase in subsequent periods, current expenditures on advertising will increase as the total advertising budget is shifted in favor of the current period which is thus made relatively cheaper. Finally, as the holding cost functions increase in either present or future periods, it results in increased advertising to generate more sales, and to reduce the terminal inventory level, thus reducing holding costs.

The results of the last three theorems are summarized in Table 1 for comparative purposes. A plus indicates that the solution is nondecreasing in the aggregate cost and a minus indicates that the solution is nonincreasing in the aggregate cost. For example, $\bar{z}^n(x)$ is nondecreasing in P_n and nonincreasing in S^n.

4. MONOTONICITY OF THE OPTIMAL POLICY IN THE PRIMARY COSTS AND THE DEMAND

In this section, we apply the results of Section 3 to obtain the monotonicity properties of the optimal policy in the costs and the basic demands. The results of this section tell us what happens to the optimal inventory level $\bar{y}(x;\cdot^n)$, the optimal controlled demand $\bar{\beta}(x;\cdot^n)$ thus to



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

TABLE 1 — *Monotonicity of the Optimal Policies in Aggregate Costs*

	Ordering-Capacity		Advertising-Revenue		Holding-Shortage Present and Future
	Present H_n	Future H^{n+1}	Present P_n	Future P^{n+1}	S^n
Inventory Level $\bar{y}(x; \cdot^n)$	−	−	−	−	−
Net Inventory Level $\bar{z}(x; \cdot^n)$	−	−	+	−	−
Advertising Level $\bar{b}(x; \cdot^n) = g^{-1}(\bar{\beta}(x; \cdot^n))$	−	+	−	+	+

the optimal advertising level $\bar{b}(x; \cdot^n)$ and to the net inventory level $\bar{z}(x; \cdot^n)$ when any primary factors change either in the current or in a future period. Note that when a factor is represented by a function, an increase in the factor means that it is now represented by a function which is greater than the previous one according to the incremental order.

Before linking the variations of the aggregate costs to those of the primary costs and thus predicting the variations of the optimal policies, we shall establish some results for the variations in the demand distribution. The set $\mathcal{F}$ of all demand distributions on the nonnegative real half line can be partially ordered in a natural way by a stochastic order defined as follows. We say $F \in \mathcal{F}$ is stochastically smaller than $G \in \mathcal{F}$, written $F \subset G$, if $F(u) \geqslant G(u)$ for all real u. Let $\mathcal{F}^n$ be the $N - n + 1$-fold cartesian product of $\mathcal{F}$. Then we say $F^n \equiv (F_n, \ldots, F_N) \in \mathcal{F}^n$ is stochastically smaller than $G^n \equiv (G_n, \ldots, G_N) \in \mathcal{F}^n$, written $F^n \subset G^n$, if $F_i \subset G_i$ for $i = n, \ldots, N$. In the sequel it is understood that $\mathcal{F}^n$ is always partially ordered by stochastic order.

THEOREM 5: For all n, $\bar{y}(x; F^n)$ and $\bar{z}(x; F^n)$ are nondecreasing in F^n on $\mathcal{F}^n$, $\bar{\beta}(x; F^n)$ and $\tilde{\beta}(x; F^n)$ are nonincreasing in F^n or $\mathcal{F}^n$; also, $C(x; F^n)$ is subadditive in (x, F_i) on $R \times \mathcal{F}_i$ for $i = n, \ldots, N$.

PROOF: We assert that $F^n \subset F'^n$ in $\mathcal{F}^n$ implies that $S^n \geqslant S'^n$ in $\mathcal{S}^n$, where S and S' are defined by equation (5). To see this, for $i \geqslant n$, let $U_i' = F'^{-1}_i(V)$ and $U_i = F_i^{-1}(V)$ where V is a uniform random variable on $[0, 1]$. By the stochastic ordering, $F_i^{-1}(V) \leqslant F'^{-1}_i(V)$ for all values of V. Thus, by convexity of S, if $\mathrm{v} < \mathrm{v}'$, then for all $z < z'$, $S(z' - \mathrm{v}) - S(z - \mathrm{v}) \geqslant S(z' - \mathrm{v}') - S(z - \mathrm{v}')$. Replacing v and v′ by $F_i^{-1}(V)$ and $F'^{-1}_i(V)$, respectively, we conclude that $S_i \geqslant S_i'$ for all i, thus $S^n \geqslant S'^n$ in $\mathcal{S}^n$. This completes the proof of the assertion. In the proof of Theorem 4, $\lambda EC(\eta(z - U); S^{n+1})$ is a function of F_n but not of S_n. However, since $C^{n+1}(\eta(x))$ is convex nondecreasing in x, $\lambda EC(\eta(x - U))$ is convex nondecreasing in x, then by Lemma A2, $\lambda EC(\eta(z - U); S^{n+1})$ is subadditive in (z, F_n). With this modification, the proof carries over implying the results.

So, as present or future expected demands increase (F stochastically smaller than G implies that the expected demand based on F is smaller than the one based on G), the optimal level of inventory should increase. Surprisingly, as present or future demand goes up, the optimal advertising level decreases. This leads us to believe that advertising is done to compensate for the lack of demand in the market, though we cannot conclude that advertising expenditures are countercyclical to market demands. This analysis will be pursued in Section 7 were fluctuations over time are investigated. But why are advertising expenditures decreasing when the demand picks up? Since, in the present model, total demand is a sum of a stochastic demand—a component that varies exogenously—and a controlled demand which depends on the advertising expenditures, advertising is not more or less effective in generating demand when the stochastic component of demand is great or small. At first, it appears that advertising expenditures should be constant, but if we think of the controlled demand as demand that is purchased at a price, its price goes up with an increase in stochastic demand. This is so, because at higher stochastic demand, holding advertising expenditures constant, more demand is generated, more inventory is needed and the capacity cost of the last unit increases with inventory, making the last unit of advertising expenditures unprofitable. Therefore, as stochastic demand goes up, advertising expenditures are curtailed.

LEMMA 2: For all n, $\bar{\beta}^n(x)$ is nondecreasing in r_n for $\eta_- < 1$.

PROOF: In equations (4) and (5), the mixed partial derivative of $P(\beta)$ with respect to β and r_n is -1, the mixed partial derivative of $S(y-\beta)$ with respect to y and r_n is $-(1-\eta_-)\bar{F}(y-\beta)$, where $\bar{F}(z) = 1 - F(z)$, and the mixed partial derivative of $S(y-\beta)$ with respect to β and r_n is $(1-\eta_-)\bar{F}(y-\beta)$. Therefore, since $S(y-\beta)$ is convex, the right hand side term in (2) is subadditive in (y, β, r_n) by example A6 and Lemma A1. By Theorem A2, the least element, $(\bar{y}(x;r^n), \bar{\beta}(x;r^n))$, minimizing the right-hand side term in (2) is nondecreasing in r_n. This completes the proof.

Using (3), (4) and (5) the variations in the primary costs result in variations of the appropriate aggregate costs and Theorems 2, 3, 4, or Table 1, Theorem 5 and Lemma 2 apply. All the results regarding variations of the optimal policies are reported in Table 2. Certain parameters have more complex variations and no result on optimal policies can be ascertained for their variations in future periods. However, by restricting ourselves to the current period, variations of the myopic policies can be linked to those of the primary costs. Again, using (3), (4) and (5) with Theorems 2, 3 and 4, variations of the myopic policies are reported in Table 3 for the primary costs when no results for the corresponding optimal policies were obtainable. The convention used in Tables 2 and 3 is identical to the one in Table 1.

TABLE 2 — *Monotonicity of the Optimal Policies in the Primary Factors*

	Demand Distribution	Capacity Cost		Advertising Cost		Holding Shortage Cost	Revenue per Unit Sold		
		Present	Future	Present	Future		r_n		
	F^n	h_n	h^{n+1}	$p(g^{-1})_n$	$p(g^{-1})^{n+1}$	s^n	$n_- < 1$	$n_- = 1$	$n_- > 1$
Inventory Level $\bar{y}(x;\cdot^n)$	+	−	−	−	−	−	+	+	±
Net Inventory Level $\bar{z}(x;\cdot^n)$	+	−	−	+	−	−	±	−	−
Advertising Level $\bar{b}(x;\cdot^n) = g^{-1}(\beta(x;\cdot^n))$	−	−	+	−	+	+	+	+	+

TABLE 3 — *Monotonicity of the Myopic Policies in the Primary Factors*

	Discount Rate	Unit Ordering Cost		Inventory Survival Rate	Demand Backlogged Rate		
		Present	Next		η_{-n}		
	λ_n	c_n	c_{n+1}	η_{+n}	$r > \lambda c$	$r = \lambda c$	$r < \lambda c$
Inventory Level $\bar{y}(x;\cdot_n)$	+	—	+	+	—	invariant	+
Net Inventory Level $\bar{z}(x;\cdot_n)$	+	—	+	+	—	invariant	+
Advertising Level $\bar{b}(x;\cdot_n) = g^{-1}(\bar{\beta}(x;\cdot_n))$	—	—	—	—	+	invariant	—

The interpretation of the results reported in Table 2 for the variations in the stochastic demand, F^n, have been discussed above following Theorem 5. The next four entries on capacity costs and advertising costs can be analyzed along the lines of their corresponding aggregate costs H and P (see Section 3). The holding-shortage cost function, s, is more interesting as it models two costs: when the terminal inventory is positive, it is the cost of holding the inventory and storing it and when the terminal inventory is negative, it is the imputed cost for having unsatisfied customers (akin to the penalty for not fulfilling the terms of a contract or, as a modeling short cut, it is a proxy to incorporate the loss resulting from a reduction in future demands due to bad service). For simplicity, let $s(z) = s_+ z^+ + s_- z^-$ where $s_+, s_- \geqslant 0$; if s' is greater than s, then $s'_+ \geqslant s_+$ and $s'_- \leqslant s_-$. So, an incrementally larger holding-shortage cost, s, means a larger holding cost, s_+, and a smaller shortage cost, s_-. With this in mind, the explanation given in Section 3 for the variations of the optimal policies in terms of S carries directly. Finally, if the sale price in the current period, r_n, increases, it is natural to increase advertising expenditures as they are more profitable and to increase inventory to close these sales quickly. This is not true when $\eta_- > 1$, this is not too surprising because, in this case, as the customers are waiting for their order, they increase the size of their order (like hoarding or panic-buying by consumers in face of shortages). Thus, by possibly having lower inventory on hand, the increased demand by unserviced customers may more than offset the penalty cost.

In Table 3, a decrease in the cost of money (an increase in the discount rate λ) fosters an increase in the inventory level as inventories become cheaper and also, because inventories are cheaper to hold, sales through advertising, i.e., sales that are purchased to a certain extent by the seller, are reduced. Finally, adjustments in the inventory levels to change in unit ordering costs for the present and the next period follow a predictable pattern: today's price goes up, postpone the purchases; tomorrow's price goes up, build up the inventory now.

5. FLUCTUATIONS OVER TIME OF THE OPTIMAL SOLUTION IN THE MYOPIC POLICIES

In this section, we will show that the fluctuations of the optimal policies over time are linked to the variations of the myopic policies. The variations of the myopic policies are taken as given, no attempt is made in this section to link them to variations of the parameters. Before studying those fluctuations, we must prove two preliminary lemmas. First, establishing a sufficient condition for the base optimal policies to equate the base myopic policies. Second, ranking the base optimal policies with respect to the base myopic policies. Let $F_n^{-1}(0)$ be the infimum of those z such that $F_n(z) > 0$.

LEMMA 3: If $\eta(z^{*n} - F_n^{-1}(0)) \leqslant y^{*n+1}$, then $y^{*n} = y_n^*$ and $\beta^{*n} = \beta_n^*$.

PROOF: Since $U \geqslant F_n^{-1}(0)$ and η is nondecreasing, $\eta(z^{*n} - U) \leqslant y^{*n+1}$ for all values of U. Thus, the point (y^{*n}, β^{*n}) is in the interior of the half-space delimited by $\eta(y - \beta - F_n^{-1}(0)) \leqslant y^{*n+1}$, and in that half-space $EC^{n+1}(\eta(y - \beta - U))$ is constant. Therefore, $y^{*n} = y_n^*$ and $\beta^{*n} = \beta_n^*$, completing the proof.

LEMMA 4: Always, $y^{*n} \leqslant y_n^*$ and $\tilde{\beta}^n(x) \geqslant \tilde{\beta}_n(x)$ for all x.

PROOF: In equation (2), both $S(y - \beta)$ and $\lambda EC^{n+1}[\eta(y - \beta - U)]$ are functions of $y - \beta$ alone. In addition, C^{n+1} is nondecreasing on R; thus $S'(z) = S(z) + \lambda EC^{n+1}[\eta(z - U)]$ is incrementally larger than $S(z)$. Hence, by Theorem 4, the result follows.

Lemmas 3 and 4 have linked the myopic policy and the optimal policy in the same period. The next three corollaries establish a relationship between optimal policies in different periods. To obtain simple results, we must define a new condition on η:

(IV) Either $y_n^* \geqslant 0$ and $\eta_{+n} \leqslant 1$ for all n, or $\eta_{+n} = \eta_{-n} = 1$ for all n.

The nonnegativity of y_n^* can be insured by Condition (III) or the restriction $y \geqslant 0$ on the minimizing space in (2).

COROLLARY 1: If Condition (IV) holds and $y_n^* \geqslant y_{n+1}^*$, then $\bar{y}^n(x) \geqslant \bar{y}^{n+1}(x)$ for all x.

PROOF: See Appendix Corollary 1'.

COROLLARY 2: If Condition (IV) holds and either $y_n^* \leqslant y_{n+1}^*$ and $y^{*n} > y^{*n+1}$, or $y_n^* < y_{n+1}^*$ and $y^{*n} \geqslant y^{*n+1}$, then $\bar{y}^{n+1}(x) \geqslant \bar{y}^{n+2}(x)$ for all x.

PROOF: See Appendix Corollary 2'.

COROLLARY 3: If Condition (IV) holds and $y^{*n} < y_n^*$, then $\bar{y}^n(x) \geqslant \bar{y}^{n+1}(x)$ for all x.

PROOF: Assume there exists x such that $\bar{y}^n(x) < \bar{y}^{n+1}(x)$. Because $\bar{y}^n(x) = y^{*n} \vee x$, $y^{*n} < y^{*n+1}$. By Lemmas 2 and 3, since $\eta(y^{*n} - F_n^{-1}(0)) < y^{*n+1}$, $y^{*n} = y_n^*$ contradicting the assumption. This completes the proof.

These three Corollaries are the building elements of our analysis of the impact on optimal policies of the variations over time in the myopic policies. Corollary 1 establishes that a decline of the myopic policies from one period to the next, implies a decline of the optimal policies. Corollary 2 states that, if myopic policies increase from one period to the next, while the optimal policies decline, then the optimal policies must also decline at a later time. Finally, from Corollary 3, we get that if an optimal policy is strictly smaller than the corresponding myopic policy, then the optimal policy in the next period is no greater than the present one. We summarize these results in a very simple theorem. To that end, we partition time into intervals T, on each of which y_n^* is unimodal in n.

THEOREM 6: If Condition (IV) holds and y_n^* is unimodal in n on an interval T, then $\bar{y}^n(x)$ is unimodal in n on T. Moreover, the least mode m' of y^{*n} on T does not exceed the least mode m of y_n^* on T. Finally, $y^{*n} = y_n^*$ for $n < m'$ on T.

PROOF: Let m be the least mode of y_n^* on T. The assertion is true if $y^{*n} = y_n^*$ for all $n \in T$. If not, there exists a least $k \in T$ such that $y^{*k} < y_k^*$. If $k > m$, Corollary 1 implies that m' equals m. If $k < m$, by Corollary 3, $y_{k+1}^* \geqslant y_k^* > y^{*k} \geqslant y^{*k+1}$. Applying Corollary 3 successively for $k, \ldots, m-1$, one sees m' equals k or $k-1$. If $k = m$, the result is immediate, completing the proof.

Theorem 6 is a natural extension of Veinott's [24] original work linking the optimal policies to the myopic policies over time in the context of a one commodity inventory model. This theorem says that when y_n^*, the myopic policies, are unimodal over time, the optimal policies y_n^* are unimodal. Moreover, the mode of y_n^* never precedes the mode of y^{*n}; when increasing, the optimal policies are equal to the myopic policies possibly with the exception of the last one; and when optimal policies decline in one period, they decline from that moment on in the interval T, i.e., until at least the moment the myopic policies increase again after having themselves declined. All these results have been expressed in terms of y_n^*, but, by Theorem 1, they extend directly to $\bar{y}^n(x) = y^{*n}\ \forall x$. The results are illustrated in Figure 1, where the graph of y_n^* is represented by the dashed line, the graph of y^{*n} by the solid line and, when there is no solid line, by the dashed line. The graph of $\bar{y}^n(x)$ is represented by the dotted line and, when there is no dotted line, by the graph of y^{*n}. One situation is not illustrated on the graph: the case when both the modes of y^{*n} and y_n^* coincide in time with possibly $y^{*n} = y_n^*$.

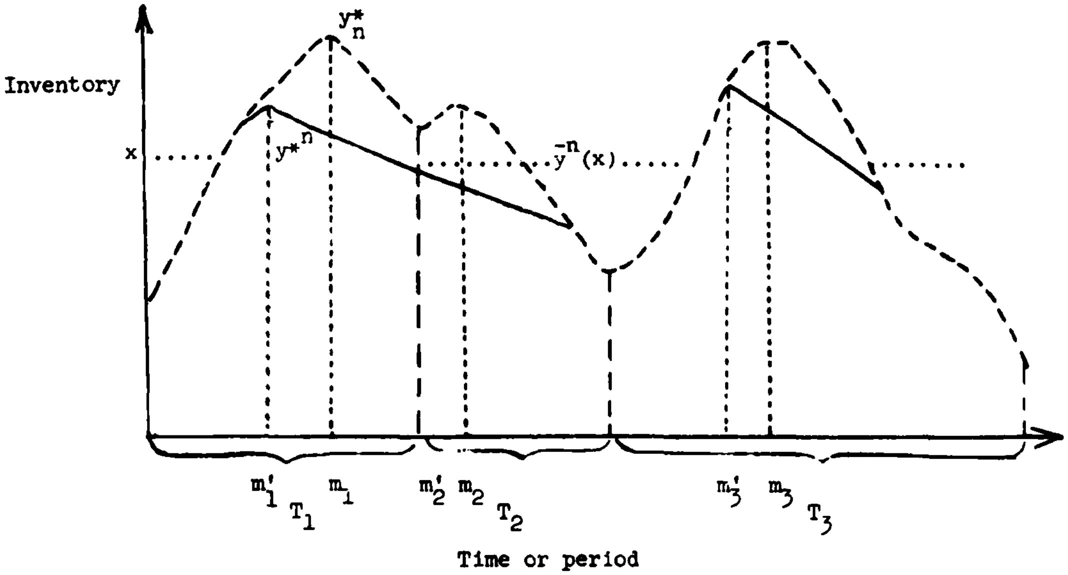


FIGURE 1. Fluctuations of the optimal inventory policies $\bar{y}^n(x)$ and base inventory level y^{*n} as a function of the variations of the myopic policies y^*_n over time.

Similar results are derived for the optimal net inventory policy which is the difference between optimal inventory policy and optimal controlled demand generated by advertising. The importance of deriving results for the net inventory level will be perceived in the next section, as they carry additional information about the fluctuations of the optimal policies over time. First, a technical lemma is proved followed by an adapted version of Corollaries 1, 2 and 3 and Theorem 6.

LEMMA 5: If $\eta_{-n} = 0$, $y^*_{n+1} = 0$, and $z^{*n} < z^{*n+1}$, then $y^{*n} = y^*_n$ and $\beta^{*n} = \beta^*_n$.

PROOF: Since $z^{*n} < z^{*n+1} = -\beta^{*n+1} \leqslant 0$, the point (y^{*n}, β^{*n}) is in the interior of the half space $\{y - \beta \leqslant 0\}$. And on that half space $EC^{n+1}(\eta_n(y - \beta - U))$ is constant because $\eta = 0$. Therefore, $y^{*n} = y^*_n$ and $\beta^{*n} = \beta^*_n$ by the convexity of the three first terms in (2) completing the proof.

COROLLARY 4: If Condition (IV) holds and $z^*_n \geqslant z^*_{n+1}$, then $z^{*n} \geqslant z^{*n+1}$.

PROOF: Assume the contrary, i.e., $z^{*n} < z^{*n+1}$. By Condition (IV), there are three possibilities, viz. (i) $\eta_{+n} = \eta_{-n} = 1$, (ii) $y^{*n+1} \geqslant 0$ and $y^{*n+1} \vee \eta_{-n} > 0$ and (iii) $y^{*n+1} = 0$ and $\eta_{-n} = 0$. In the first two cases $\eta_n(z^{*n}) < y^{*n+1}$ so $z^{*n} = z^*_n$ by Lemma 3. In the third case $z^{*n} = z^*_n$ by Lemma 5. Thus, by Lemma 4, in all cases $z^{*n} = z^*_n \geqslant z^*_{n+1} \geqslant z^{*n+1}$, which is a contradiction and completes the proof.

COROLLARY 5: If Condition (IV) holds and $z^{*n} < z^*_n$, then $z^{*n} \geqslant z^{*n+1}$.

PROOF: Assume the contrary, i.e., $z^{*n} < z^{*n+1}$. It follows from Condition (IV), as in the proof of Corollary 4, that $z^{*n} = z^*_n$, which is a contradiction and completes the proof.

COROLLARY 6: If Condition (IV) holds and if either $z^*_n \leqslant z^*_{n+1}$ and $z^{*n} > z^{*n+1}$ or $z^*_n < z^*_{n+1}$ and $z^{*n} \geqslant z^{*n+1}$, then $z^{*n+1} \geqslant z^{*n+2}$.

PROOF: Assume the contrary, i.e., $z^{*n+1} < z^{*n+2}$. Then from Condition (IV), as in the proof of Corollary 4, $z^{*n+1} = z^*_{n+1}$. Hence, by Lemma 4, $z^{*n+1} = z^*_{n+1} \geqslant z^*_n \geqslant z^{*n} \geqslant z^{*n+1}$ with either the first or third inequality being strict. This is a contradiction and completes the proof.

THEOREM 7: If Condition (IV) holds and z^*_n is unimodal in n on T, then z^{*n} is unimodal in n on T. Moreover, the least mode m'' of z^{*n} on T does not exceed the least mode of z^*_n on T. Finally, $z^{*n} = z^*_n$ for $n < m''$ on T.

PROOF: Identical to that of Theorem 6 with Corollaries 4 and 6 replacing 1 and 3.

We can now define a new condition to assure strict results in Corollaries 1, 2, 3, 4, 5, and 6:

(V) If $\eta_{+n} = 1$, then $F_n^{-1}(0) > 0$.

In all the results concerning the base inventory, y^{*n}, in Corollaries 1, 2 and 3 and the base net inventory, z^{*n} in Corollaries 4, 5 and 6, the last inequality sign can be replaced by a strict one, under Condition (V). This is so because, if either $\eta_{+n} < 1$ or $\eta_{+n} = 1$ and $F_n^{-1}(0) > 0$, $z^{*n} \leqslant z^{*n+1}$ implies $\eta(z^{*n} - U) < y^{*n+1}$.

6. FLUCTUATIONS OVER TIME OF THE OPTIMAL SOLUTION IN THE AGGREGATE COSTS

In this section, we link the fluctuations of the optimal solutions with the variations over time of one aggregate cost while the other two remain stationary. Thus, we can tell, for example, when the advertising costs change with the seasons, what should be the pattern over time of the optimal policies without any computations. Clearly, without computations, we cannot hope to find the exact value of the optimal policies, but many potential candidates can be discarded at a glance. Say, optimal advertising expenditures must peak during the cheapest month for advertising, so any planned expenditures over time which do not peak at that time, are not optimal and should not be implemented with that timing.

For notational simplicity, let $y^*(\omega_n)$ denote the myopic base inventory in period n when ω varies from one period to next; as usual, ω can take the value H, P or S. This obviously applies to all other policies as well. The reader should note that, for the first time, stationarity over time of certain costs is required. By convention, only the parameters specifically exhibited vary over time. Also, the reader should be aware that comparing $y^*(\omega^n)$ and $y^*(\omega^{n+1})$ is not that simple even when ω_n is comparable to ω_{n+1} (larger or smaller by the incremental ordering), because we are comparing $(\omega_n, \omega_{n+1}, \omega_{n+2}, \ldots)$ with $(\omega_{n+1}, \omega_{n+2}, \ldots)$, not $(\omega_n, \omega_{n+1}, \omega_{n+2}, \ldots)$ with $(\omega_n', \omega_{n+1}', \omega_{n+2}', \ldots)$ as was done in Sections 3 and 4; the former are not comparable, while the latter were. However, by Theorems 2, 3 and 4, we can derive results on the myopic base stocks because we compare the period one by one, i.e., compare ω_n to ω_{n+1}, not ω^n to ω^{n+1}. Since all other parameters are stationary, comparing ω_n to ω_{n+1} is equivalent to comparing ω_n with ω_n' in the same period. Therefore, if $\omega_n \leqslant \omega_{n+1}$, and $y^*(\omega^n)$ is increasing in ω_n, then $y^*(\omega_n)$ is increasing in ω_n and $y^*(\omega_n) \leqslant y^*(\omega_{n+1})$. Expanding on this theme, let ω_n, $n = 1, \ldots, N$ be such that ω_n is comparable to ω_{n+1} for all n. It follows that all the $y^*(\omega_n)$ are comparable with the same pattern. Note that ω_n and ω_{n+2} are not necessarily comparable because possibly $\omega_n \leqslant \omega_{n+1}$ and $\omega_{n+1} > \omega_{n+2}$. Thus, we can replace the conditions on the myopic base stocks in Theorems 6 and 7 by a condition on the parameters directly and the conclusions of these theorems still hold.

The rest of this section is concerned with establishing the relationship over time of the fluctuations of the optimal policies and of the base stocks with the variations of the aggregate costs. Corresponding to the variations of each of these aggregate costs, there are types of fluctuations for the optimal policies labeled Type P, H and S in the obvious fashion. This convention will prove useful when primary costs are examined in the next section.

THEOREM 8: (Type P). Let Condition (IV) hold. If the model is stationary, except possibly for the P_n, and if P_n and P_{n+1} are comparable for all n, the myopic policy is optimal in every period for all intitial inventories $x \leqslant y^*(P_1)$. Moreover, $(y^*(P^n), \beta^*(P^n), z^*(P^n)) = (y^*(P_n), \beta^*(P_n), z^*(P_n))$.

PROOF: See Appendix.

If $x \leqslant y^*(P_1)$, the optimal policy $(\bar{y}^n(x), \bar{\beta}^n(x))$ is equal to the base myopic policy (y_n^*, β_n^*). It says that when advertising costs vary over time, the manager should react as if short-sighted, i.e., pay no attention to the parameters of future periods. Also, both advertising expenditures and inventory levels are countercyclical to variations in the advertising cost functions, with perfectly matching cycles. Noting that the net inventory positions are procyclical indicates that the induced fluctuations over time are larger for the advertising expenditures than for the inventory. These results are what we would have expected, at least regarding the anticyclical behavior of advertising expenditures and inventory positions. The results of this theorem are illustrated in Figure 2, where $y^*(P^n)$, $\beta^*(P^n)$, and $z^*(P^n)$ are plotted. The solid line represents $y^*(P^n)$, the dashed-dotted line $\beta^*(P^n)$ and the dotted line $z^*(P^n)$. The reader should not forget that the driving force behind these fluctuations are the variations of P_n, which are not represented on the graph—they are on an uncomparable scale—and which are perfectly anticyclical to $y^*(P_n) = y^*(P^n)$. For continuity of the representation of the optimal solutions, the graphs of $y^*(P^n)$ in Figure 2 and y_n^* in Figure 1 are identical.

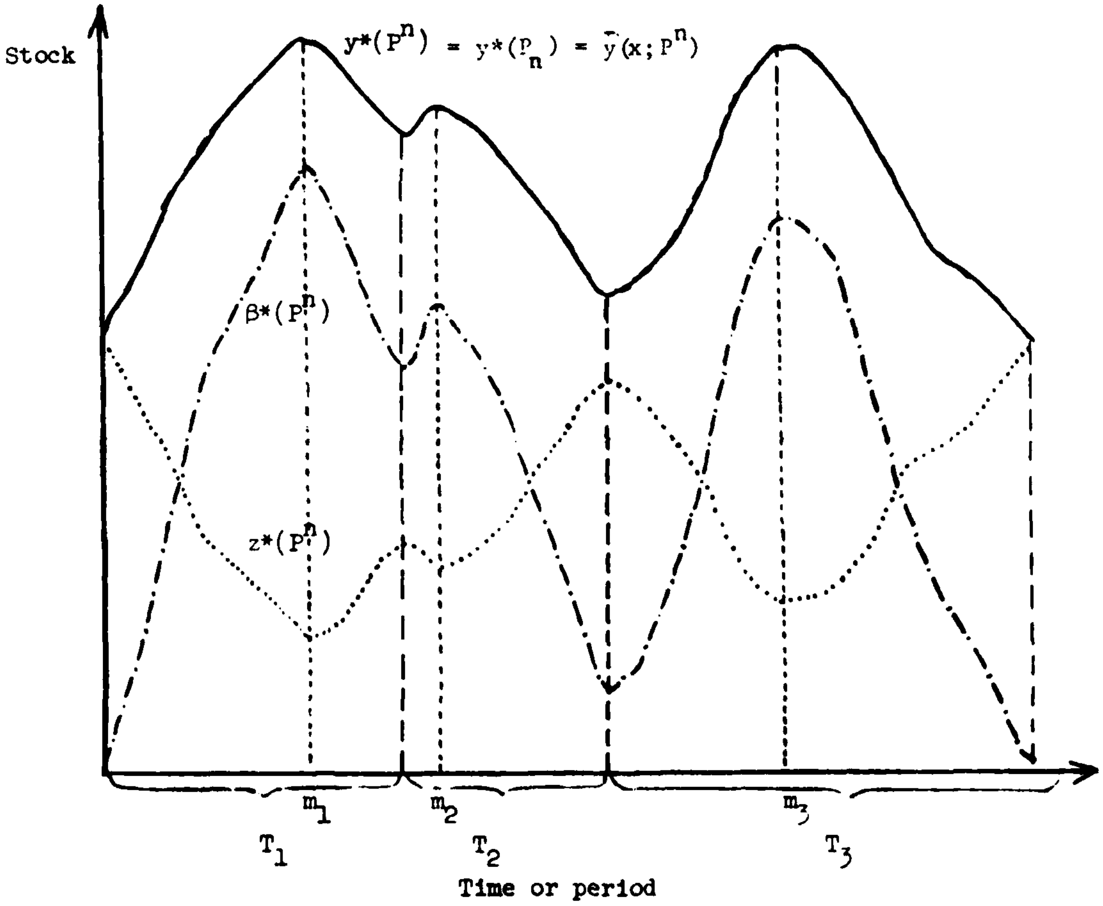


FIGURE 2. Fluctuations (Type P) of the base stocks with respect to variations in P_n over time, whose variations are perfectly anticyclical to those of $y^*(P_n)$.

Next, we examine the fluctuations (Type H) induced by the variations of the aggregate ordering-capacity cost.

THEOREM 9: (Type *H*). Let Condition (IV) hold. If the model is stationary, except possibly for the H_n, and if H_n in $\mathscr{H}^*$ is unimodal in n on an interval T with least mode m, then $y^*(H^n)$ and $z^*(H^n)$ are unimodal in n on T with respective least modes m' and m'', such that $m'' \leqslant m' \leqslant m$. Moreover, $(y^*(H^n), \beta^*(H^n), z^*(H^n)) = (y^*(H_n), \beta^*(H_n), z^*(H_n))$ for $n \leqslant m''$, and $\beta^*(H^n)$ is nondecreasing in $n \leqslant m$.

PROOF: See Appendix.

As expected from the discussion following Theorem 2, the base levels y^*, β^* and z^*, vary in an anticyclical manner to the variations of the aggregate ordering-capacity cost overtime. The time of highest cost corresponds to the time of lowest base levels, but the base inventory and net inventory peak before the moment when the aggregate cost reaches its lowest level. Since inventory is not necessarily sold immediately, the inventory purchases of yesterday may last until tomorrow when the costs start rising again; so, in anticipation, inventory levels are curtailed while the costs are still declining. On the contrary, advertising expenditures which generate current sales do not peak before the time of lowest cost. The results of Theorem 9 are illustrated in Figure 3, whence the convention used in Figure 2 is maintained; in addition $y^*(H_n)$ is pictured by a dashed line, and the variations of H_n not pictured in Figure 3 are exactly anticyclical to those of $y^*(H_n)$.

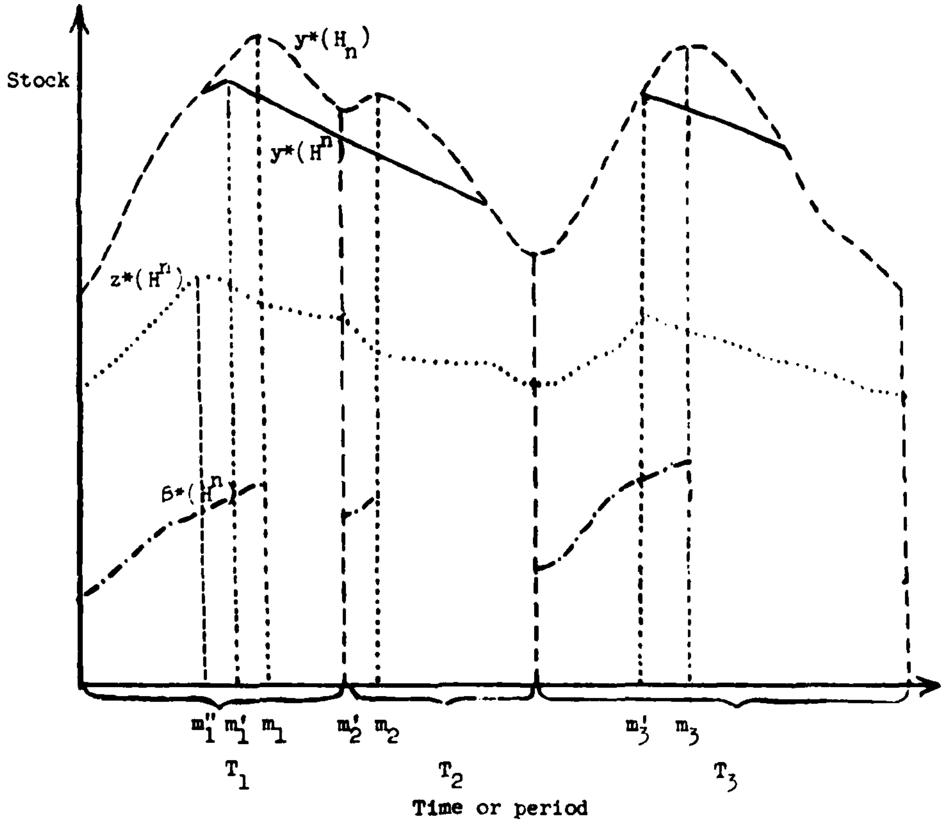


FIGURE 3. Fluctuations (Type H) of the base stocks with respect to the variations in H_n over time, whose variations are perfectly anticyclical to those of $y^*(H_n)$.

Finally, we turn our attention to the fluctuations (Type S) induced by the variations over time in the aggregate shortage holding costs, S_n.

THEOREM 10: (Type S). Let Condition (IV) hold. If the model is stationary, except possibly for the S_n, and if S_n in $\mathcal{S}^*$ is unimodal in n on an interval T, then $y^*(S^n)$, $z^*(S^n)$ and $\beta^*(S^n)$ are unimodal in n on T. Also, the least mode m' of $z^*(S^n)$ is no larger than that of S_n. Moreover, each mode of $z^*(S^n)$ is no larger than that of S_n. Moreover, each mode of $z^*(S^n)$ is a mode of $y^*(S^n)$ and $-\beta^*(S^n)$, and $(y^*(S^n), \beta^*(S^n), z^*(S^n)) = (y^*(S_n), \beta^*(S_n), z^*(S_n))$ for $n < m'$.

PROOF: See Appendix.

Again, as expected from the discussion following Theorem 4, the base levels of y^*, $-\beta^*$ and z^* vary in an anticyclical manner to the variations of the aggregate shortage holding cost over time. The time of highest cost corresponds to the moment of lowest inventory and net inventory positions and highest advertising expenditures. For the same reason as in the previous theorem, the inventory peaks before the time of lowest cost, in anticipation of future higher costs. This particular type will be discussed further in the next section in conjunction with variations in the demand. These results are illustrated in Figure 4 where the conventions used in Figure 3 are maintained.

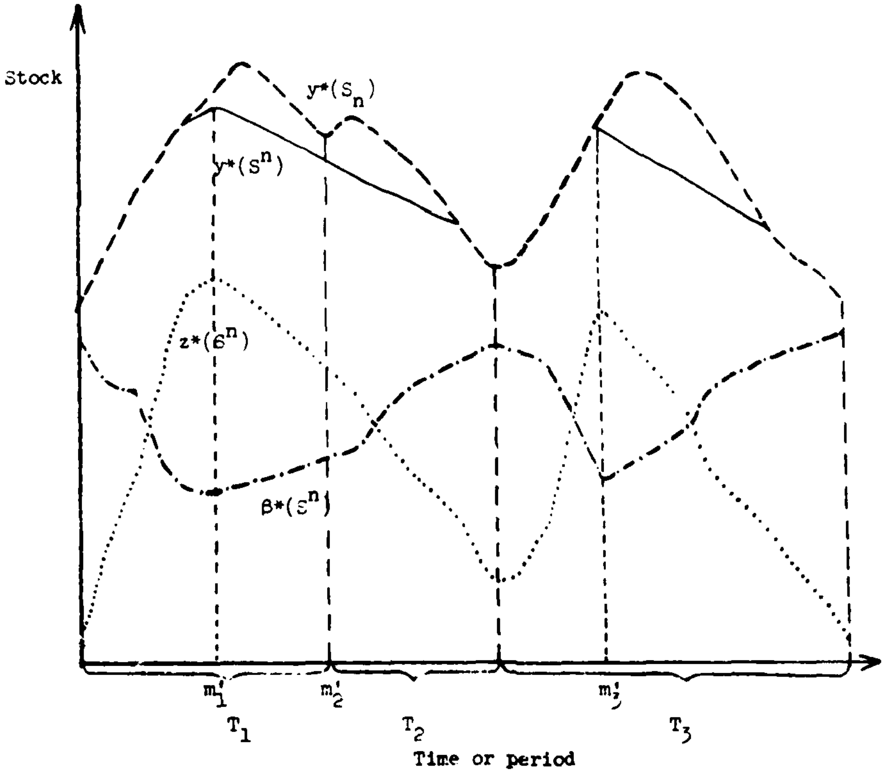


FIGURE 4. Fluctuations (Type S) of the base stocks with respect to variations in S_n over time, whose variations are perfectly anticyclical to those of $y^*(S_n)$.

As mentioned earlier, results of Theorems 8, 9 and 10 extend directly from the base inventory levels y^{*n} to optimal inventory levels $\bar{y}^n(x)$ because $y^n(x) = y^{*n} \vee x$. They extend also to $\bar{\beta}^n(x)$ and $\bar{z}^n(x)$ when the myopic policies are optimal using Theorems 2, 3 and 4. But, when myopic policies are not optimal (see Theorems 9 and 10), they cannot be extended to $\bar{\beta}^n(x)$ and $\bar{z}^n(x)$, because in Lemma 3, $\beta_n^* = \beta^{*n}$ does not imply $\bar{\beta}_n(x) = \bar{\beta}^n(x)$; however, the fluctuations of $\bar{\beta}^n(x)$ and $\bar{z}^n(x)$ are unlikely to deviate strongly from β^{*n} and z^{*n}, though a counter example should not be hard to construct.

7. FLUCTUATIONS OVER TIME OF THE OPTIMAL SOLUTION IN THE PRIMARY COSTS AND IN THE BASIC DEMAND

In this section, we will establish that, under variations of either the primary costs or the basic demand, the fluctuations of the optimal solution are of either Type *H*, *P*, or *S*, as defined by Theorems 8, 9 and 10. Since Condition (IV) holds in every theorem, we will assume throughout this section that Condition (IV) holds.

Also for clarity, the results, which are obtained directly from Tables 2 and 3, will be summarized in Table 4. The first column of the table contains the nonstationary factor in each case; once more, we remind the reader that only one factor in each case is nonstationary over time. The second column indicates the ordering used. In the third column, a minus shows that the order in the set is reversed, and a plus is used otherwise. Note that in Theorems 8, 9 and 10 the order is reversed. In the fourth column, the types of fluctuations are indicated by either *H*, *P* or *S*, as defined in the previous section.

TABLE 4 — *Fluctuations of the Optimal Policies with the Primary Costs and Demand*

Nonstationary parameter	Ordering		
	Type	Natural (+) Reversed (−)	Fluctuation Type
Advertising cost $p(b^{-1})$	incremental	−	P
Sales price r, $(\eta_- = 1)$	real	+	P
Capacity cost h	incremental	−	H
Backlogged demand factor η_-, $(r = \lambda c)$	real	+,−	invariant
Backlogged demand factor η_-, $(r < \lambda c)$	real	+	S
Backlogged demand factor η_-, $(r > \lambda c)$	real	−	S
Inventory preserving factor η_+	real	+	S
Discount factor λ	real	+	S
Holding and shortage cost s	incremental	−	S
Basic demand distribution F	stochastic	+	S

Most of the results in Table 4 can be interpreted along the lines that followed Theorems 8, 9 and 10 and the reader is referred back to the previous section. There are, however, three cases that retain our attention: variations in selling price, in the discount rate (cost of financing the operation) and in the stochastic demand. The strongest results are obtained for the variations in selling price when unsatisfied customers wait till their order is serviced. In that case, the optimal inventory level and the advertising expenditures move in perfect harmony with the variations in price, while the net inventory position varies in the opposite direction implying that the swings of the advertising expenditures are relatively greater than those of the inventory position. When the demand varies over time, the inventory level and the net inventory fluctuate in the same fashion while the advertising expenditures are countercyclical. In this case, the peaks of inventory precede those of the demand; as sales are uncertain, no one wants to have a too large inventory on hand during the decline in demand following the peak. The cautious (and correct) approach is to retrench early. There is no need for this cautious approach when prices vary, because in that case the demand is stable, and the objective is to sell at the right time and, at no time is it better than when the prices peak. In a more general and realistic model, both demand and prices vary simultaneously, likely in harmony. Our results state clearly that the inventory level should follow that pattern, but they are ambiguous regarding advertising expenditures. Finally, when the discount rate increases or, equivalently, when the interest rate declines, the financial cost of holding the inventory (imbedded in this model as the discounted value of the terminal inventory position) is reduced and it becomes cheaper to hold larger inventory. As expected, advertising expenditures rise as the interest rate rises to help liquidate the inventory position which has become costlier.

One case is not covered in this table, namely the variations of r over time when $\eta_- < 1$. Let r_n be unimodal in n on T. Thus, under Condition (IV) and by Table 2, $y^*(r^n)$ is unimodal in n on T with least mode m not exceeding the least mode of r_n. Also, by Theorem 6, $\beta^*(r^{n-1}) = \beta^*(r_{n-1}) \leqslant \beta^*(r_n) = \beta^*(r^n)$ for $n < m$. In addition, $\beta^*(r^{m-1}) = \beta^*(r_{m-1}) \leqslant \beta^*(r_m) \leqslant \beta^*(r^m)$, by Lemma 4. Moreover, if $y^*(r^n) \geqslant y^*(r^{n+1})$ and $r_n \leqslant r_{n+1}$, $z^*(r^n) \geqslant \eta(z^*(r^n) - F_n^{-1}(0)) \geqslant y^*(r^{n+1}) \geqslant z^*(r^{n+1})$ by Lemma 3; thus, $\beta^*(r^n) = \hat{\beta}(z^*(r^n); r_n)$ is nonincreasing in $z^*(r^n)$ and $-r_n$, by an argument associated with equation (15). Summarizing $\beta^*(r^n) \leqslant \beta^*(r^{n+1})$ when $r_n \leqslant r_{n+1}$. So we increase the inventory position when the sale price increases, but we start to reduce it before prices peak. On the other hand, as long as prices rise, advertising expenditures rise, as their effect results in an immediate sale.

Obsolescence Probabilities

The discount factors λ_n with $0 \leqslant \lambda_n \leqslant 1$ can be interpretd as the probability that the process will continue in period $n + 1$ given that the process is active in period n. Thus, $1 - \lambda_n$ is the obsolescence probability in period $n+1$. Pierskalla [10] established that, with nondecreasing obsolescence probabilities, i.e., $\lambda_n \geqslant \lambda_{n+1}$ for all n and a positive ordering cost, the base stocks are such that $y^{*1} \geqslant \ldots \geqslant y^{*N}$. In Table 4, we obtain a more general result concerning the relationship between the obsolescence probabilities and the base stocks. For example, if λ_n is unimodal in n, then y^{*n} is unimodal in n with the time of its least peak not exceeding that of λ_n. If λ_n is nonincreasing in n, we obtain Pierskalla's result that the same is so of y^{*n}.

8. EXTENSIONS

In this section, we generalize the model in two separate directions: convex ordering cost for the inventory and common time lag in delivery and advertising.

First, if we replace the previous inventory ordering cost by the convex ordering cost $\tilde{c}(z) = c(z) + cz$ in (2), we obtain:

THEOREM 14: If η is linear, the assertions of Tables 1, 2 and 3 continue to hold for the model of this section.

PROOF: Upon noticing that the additional term $c(y - x)$ in (2) is independent of H, P and S, the proofs of Theorems 2, 3 and 4 can be used directly since η is linear. Thus, Tables 1, 2 and 3 hold true in this case, completing the proof.

Karlin [5] claimed that, with convex ordering cost and no advertising, i.e., $\bar{\beta}(x) \equiv 0$, the fluctuations of the optimal inventory policy can be linked to the variations of the demand distribution over time. Upon reflection, this seems unlikely as the optimal policy responds not only to variations in demand but also must smooth the size of the orders from one period to the next, due to the added constraint of convex ordering cost. The following example provides a counterexample to an analogue of Corollary 1 for convex ordering cost and no advertising. It will show that even if the myopic policy falls and rises, the optimal policy keeps rising to smooth the ordering cost overtime. Without such an analogue corollary, it is not possible to establish a theorem analogue to Theorem 6, and therefore the fluctuations of the optimal policy cannot be predicted from those of the myopic policy.

EXAMPLE: Let $\eta = 1$, $H = 0$, $P = 0$, $c(z) = S(z) = z^2/2$, $U^1 = (10, 0, 100)$ and $N = 3$. The myopic policies in each period are $\bar{y}_1(x; 10) = ((10 + x)/2) \vee x$, $\bar{y}_2(x; 0) = (x/2) \vee x$ and $\bar{y}_3(x; 100) = ((100 + x)/2) \vee x$. The optimal policies in each period are: $\bar{y}^1(x; 10, 0, 100) = ((5x + 180)/13) \vee x$, $\bar{y}^2(x; 0, 100) = ((2x + 100)/5) \vee x$ and $\bar{y}^3(x; 100) = ((100 + x)/2) \vee x$. Therefore, when $x = 0$, $\bar{y}_1(0;10) = 5 > \bar{y}_2(0;0) = 0 < \bar{y}_3(0; 100) = 50$, but $\bar{y}^1(0;10, 0, 100) = 13\ 11/13 < \bar{y}^2(0;0, 100) = 20 < \bar{y}^3(0;100) = 50$.

Second, if a common interval of time, ν, is allowed to occur between the moments of product ordering and of advertising purchase and the moments of the delivery of the product and of the advertising effect, all the results of the previous sections hold, provided that $\eta_+ = \eta_- = 1$ and (5) is replaced by

$$S(z) = Es(z - V) - \lambda c'z,$$

where V is a random variable whose distribution is the convolution of $F_n, \ldots, F_{n+\nu}$. In particular, if the comparison criteria $F_n \subset F_{n+1}$ or $F_n \supset F_{n+1}$ is replaced by $F_n \subset F_{n+\nu+1}$ or $F_n \supset F_{n+\nu+1}$, as in Veinott [17], the results of Table 4 apply to the fluctuations of the basic demand distribution.

9. CONCLUSIONS

This paper is an example of the use of dynamic programming in a nonstationary environment where properties of the optimal policies were derived successfully without resorting to numerical computations or severe limitations on the cost functions—such as restricting them to be quadratic. The methodology used in this paper can be applied directly to discrete time control problems, if the controls and nonstationary parameters of interest exhibit some subadditivity properties.

To come back to the specific model of inventory-advertisement treated in this paper, its most peculiar feature is the fact that advertising expenditures should take place during periods of low demand so as to smooth total demand over time. This corresponds well to our notion of a sales campaign during the low season (cars in January, appliances around the same time, winter equipment around spring time, etc.). However, it does not capture the fact that often advertising expenditures peaked just prior to or at the peak of the season (ice tea commercials on TV during the summer, etc.). A more complex demand function is needed to handle this problem and possibly the simultaneous presence of two types of advertisement—procyclical and countercyclical. Another factor that is neglected in this paper is the simultaneous presence of more than one firm in the market, firms which compete for many of the same customers. This gaming factor could very well reverse the results obtained here, but I would not like to venture a guess at this stage.

ACKNOWLEDGMENTS

The author thanks Arthur F. Veinott, Jr., who, as a teacher and thesis director, has devoted countless hours helping to refine many ideas. He is particularly grateful for the privilege of using his unpublished work on lattice programming and for suggesting the problem posed by fluctuations of demand on the choice of optimal policies. He also gratefully acknowledges the financial support from the Conseil des Arts du Canada, the Conseil National de Recherches du Canada and the Office of Naval Research. The author owes a great deal to the referee for his help and guidance in rendering this paper more intuitive and mathematically readable

REFERENCES

[1] Arrow, K.J. and M. Nerlove, "Optimal Adverstising Policy under Dynamic Conditions," Economica, *29*, 129-142 (1962).

[2] Balcer, Y., "Optimal Advertising and Inventory Policy with Random Demands," Unpublished Ph.D. Thesis, Operations Research, Stanford University, Stanford, CA (1974).

[3] Balcer, Y., "Partially Controlled Demand and Inventory Control: An Additive Model," Naval Research Logistics Quarterly, *27*, 273-288 (1980).

[4] Clement, W.E., P.L. Henderson and C.P. Eley, "The Effects of Different Levels of Promotional Expenditure on Sales of Fluid Milk," Economic Research Service, USDA, Washington, DC (1965).

[5] Karlin, S., "Dynamic Inventory Policy with Varying Stochastic Demand," Management Science, *6*, 231-258 (1960).

[6] Kuehn, A., "A Model for Budgeting Advertising," in *Mathematical Models and Methods in Marketing*, Bass, F.M., Editor (Richard D. Irwin, Inc. 1961) Homewood, IL.

[7] Kuehn, A., "How Advertising Performance Depends on Other Marketing Factors," Journal of Advertising Research, *2*, 2-10 (1962).

[8] Miercort, F.A., "Some Effects of Advertising and Prices on Optimal Inventory Policy," Department of Operations Research, TR 104, Stanford University, Stanford, CA (1968).

[9] Palda, K.S., "The Measurement of Cumulative Advertising Effects," Journal of Business, *38*, 162-179 (1965).

[10] Pierskalla, W., "Inventory Problem with Obsolescence," Naval Research Logistics Quarterly, *16*, 218-228 (1969).

[11] Shryer, W.A., "*Analytical Advertising*," Business Service Corporation, Detroit, MI (1912).

[12] Simon, J.L., "The Effect of Advertising on Liquor Brand Sales," Journal of Marketing Research, *6*, 301-313 (1969).

[13] Simon, J.L., *Issues in the Economics of Advertising* (University of Illinois Press, Urbana, IL 1970).

[14] Stone, R.F., *Successful Direct Mail Advertising and Selling* (Prentice-Hall, New York, 1955).

[15] Telser, L.G., "Advertising and Cigarettes," Journal of Political Economy, *70*, 471-499 (1962).

[16] Topkis, D.M., private communication (1969).

[17] Topkis, D.M., "Minimizing a Submodular Function on a Lattice," Operations Research, *26*, 305-321 (1978).

[18] Topkis, D.M. and A.F. Veinott, Jr., "Monotone Solutions of External Problem on Lattices" (abstract), Eighth International Symposium on Mathematical Programming, Stanford University, Stanford, CA, 131 (1973).

[19] Veinott, A.F., Jr., "Optimal Stockage Policies with Nonstationary Stochastic Demand," in *Multistage Inventory Models and Techniques*, H.E. Scarf, D.M. Gilford and M.W. Shelly, Editors (Stanford University Press, Stanford, CA, 85-115, 1963).

[20] Veinott, A.F., Jr., "Optimal Policy for a Multiproduct, Dynamic, Nonstationary Inventory Problem," Management Science, *12*, 202-222 (1963).

[21] Veinott, A.F., Jr., "Inventory and Production Control," lecture notes, unpublished (1968).

[22] Veinott, A.F., Jr., private communication (1969).

[23] Veinott, A.F., Jr., "Inventory and Production Control," lecture notes, unpublished (1970).

[24] Veinott, A.F., Jr., "Subadditive Functions on Lattices in Inventory Theory," (submitted for publication).

[25] Veinott, A.F., Jr., "Monotone Solutions of External Problems on Lattices" (submitted for publication).

APPENDIX

THEOREM 2:

PROOF: Since $S(z) + \lambda EC[\eta(z - U); H^{n+1}]$ is convex in z, the term inside the braces in (2) is *subadditive in* (y, β, H_n) *on* $R \times R_+ \times \mathscr{H}'^{*}_{n}$ *by examples A7 and A8 and Lemma 1.* Therefore, by Theorem A2, the least element, $(\bar{y}(x; H^n), \bar{\beta}(x, H^n))$, minimizing the right hand side of (2) over the sublattice $\{y \geqslant x; \beta \geqslant 0\}$ is nondecreasing in H_n on $\mathscr{H}^{*}_{n}$. Letting the sublattice be $\{y = x, \beta \geqslant 0\}$, we obtain directly that $\bar{\beta}(x; H_n)$ is independent of H_n.

On eliminating β by making the change of variables, $z = y - \beta$, recursion (2) becomes

$$C(x; H^n) = \min_{y \geqslant x, y \geqslant z} \{H_n(y) + P(y - z) + S(z) + \lambda EC[\eta(z - U); H^{n+1}]\}. \tag{6}$$

We now show inductively that $C(x; H^n)$ is subadditive in (x, H^n) on $R \times \mathscr{H}^{n*}$. Since $C(x; H^{N+1}) \equiv 0$, the result holds trivially for $C(x; H^{N+1})$. By the induction hypothesis, example A5, and the fact η is nondecreasing, $\lambda EC[\eta(z - U); H^{n+1}]$ is subadditive in (z, H^{n+1}) on $R \times \mathscr{H}'^{n+1*}$. Since P is convex, by examples A6 and A8 and Lemma 1, the term inside the braces in (14) is subadditive in (y, z, H^n) on $R^2 \times \mathscr{H}'^{n*}$. By Theorem A1, $C(x; H^n)$ is subadditive in (x, H^n) on $R \times \mathscr{H}^{n*}$, terminating the induction. By Theorem A2, the 2-lexicographicaly least element, $(\bar{y}(x; H^n), \bar{z}(x; H^n))$, is nonincreasing in H^n on $\mathscr{H}^n$.

If we make the same change of variables used above but instead eliminate y, recursion (2) becomes

$$C(x; H^n) = \min_{\beta - (-z) \geqslant x, \beta \geqslant 0} \{H_n(\beta - (-z)) + P(\beta) + S(-(-z)) + \lambda EC[\eta(-(-z) - U); H^{n+1}]\}. \tag{7}$$

The term inside the braces in (7) is subadditive in $(-z,\beta,H^{n+1})$ on $R \times R_+ \times \mathscr{H}'_i$ for $i > n$ by examples A6 and A8 because H_n is convex. By Theorem A2, the least element, $\bar{\beta}(x;H^{n+1})$ right hand side of (7) is nondecreasing in H^{n+1} on $\mathscr{H}^n$. Let $y = x$, so the right hand side of (2) becomes

(8) $$\min_{\beta \geqslant 0} \{H_n(x) + P(\beta) + S(x-\beta) + \lambda EC[\eta(x-\beta-U); H^{n+1}]\}$$

The term inside the braces in (8) is subadditive in (β,H_i) on $R_+ \times \mathscr{H}'_i$ for $i > n$. By Theorem A2, the least element, $\tilde{\beta}(x;H^{n+1})$ minimizing (8) over $\{\beta \geqslant 0\}$ is nondecreasing in H^{n+1} on $\mathscr{H}^n$, completing the proof.

The proofs of Theorems 3 and 4 are sketchy, as they follow the line of argument used in Theorem 2. For detailed proofs, see Balcer [2].

THEOREM 3.

PROOF: On eliminating β by making the change of variables, $z = y - \beta$, recursion (2) becomes

(9) $$C(x;P^n) = \min_{y \geqslant x, y \geqslant z} \{H(y) + P_n(y-z) + S(z) + \lambda EC[\eta(z-U); P^{n+1}]\}.$$

By induction $C(x;P^n)$ is subadditive in (x,P_i) on $R \times \mathscr{P}'^*_i$ for $i = n, \dots, N$. By Theorem A2 and equation (9), the 2-lexicographically least element, $(\bar{y}(x;P^n), \bar{z}(x;P^n))$, minimizing the right hand side of (9) over $\{y \geqslant x,\ y \geqslant z\}$, is nonincreasing in P^{n+1} on $\mathscr{P}^n$. By Theorem A2 and equation (2), the least element, $(\bar{y}(x;P^n), \bar{\beta}(x;P^n))$, minimizing the right hand side of (2) over $\{y \geqslant x,\ \beta \geqslant 0\}$, is nonincreasing in P_n on $\mathscr{P}_n$.

If we make the same change of variables used above but instead eliminate y, the right hand side of recursion (2) becomes

(10) $$\min_{\beta-(-z) \geqslant x, \beta \geqslant 0} \{H(\beta-(-z)) + P_n(\beta) + S(-(-z)) + \lambda EC[\eta(-(-z)-U); P^{n+1}]\}$$

By Theorem A2, the 2-lexicographically least element, $(\bar{\beta}(x;P^n), -\bar{z}(x;P^n))$, minimizing the term inside the braces in (10) is nonincreasing in P_n and nondecreasing in P^{n+1} on $\mathscr{P}^n$.

In equation (2), let $y = x$ so its right-hand side becomes

(11) $$\min_{\beta \geqslant 0} \{H(x) + P_n(\beta) + S(x-\beta) + \lambda EC[\eta(x-\beta-U); P^{n+1}]\}$$

Thus, by Theorem A2, the least element, $\tilde{\beta}(x;P^n)$, is nondecreasing in P^{n+1} and nonincreasing on P_n on $\mathscr{P}^n$. This completes the proof.

THEOREM 4.

PROOF: First, on making the change of variables $\beta = y - z$, recursion (2) becomes

(12) $$C(x;S^n) = \min_{y \geqslant x, y \geqslant z} \{H(y) + P(y-z) + S_n(z) + \lambda EC[\eta(z-U): S^{n+1}]\}$$

By induction that $C(x;S^n)$ is subadditive in (x,S^n) on $R \times \mathscr{S}'^{n*}$. By Theorem A2, the 2-lexicographically least element, $(\bar{y}(x;S^n), \bar{z}(x;S^n))$, of the minimizing set $M^n(x)$ is nonincreasing in S^n on $\mathscr{S}^n$.

If instead of eliminating β, we eliminate y by the change of variables $y - \beta = z$, recursion (2) becomes

$$C(x;S^n) = \min_{\beta-(-z)\geqslant x,\beta\geqslant 0} \{H(\beta - (-z)) + P(\beta) + S_n(-(-z)) + \lambda EC[\eta(-(-z) - U); S^{n+1}]\}. \tag{13}$$

By Theorem A2, the 2-lexicographically least element, $(\bar{\beta}(x;S^n), -\bar{z}(x;S^n))$, minimizing the term in braces in the right hand side of (13) is nonincreasing in S^n on $\mathcal{S}^n$.

If we set $y = x$, the right hand side of (2) becomes

$$\min_{\beta\geqslant 0} \{H(x) + P(\beta) + S_n(x - \beta) + \lambda EC[\eta(x - \beta - U);S^{n+1}]\}. \tag{14}$$

By Theorem A2, the least element, $\tilde{\beta}(x;S^n)$, minimizing the term inside the braces in (14), is nondecreasing in S^n on $\mathcal{S}^n$, completing the proof.

Let $\bar{\eta}(x) = \eta(x) \vee x$, so $\bar{\eta}(x) = (\eta_+ \vee 1)x^+ - (\eta_- \wedge 1)x^-$.

COROLLARY 1: If $y_n^* \geqslant y_{n+1}^*$, then $\bar{\eta}(\bar{y}^n(x)) \geqslant \bar{y}^{n+1}(x)$ for all x.

PROOF: Suppose there is an x such that $\bar{\eta}(\bar{y}^n(x)) < \bar{y}^{n+1}(x)$. We must have $x < y^{*n} \vee y^{*n+1}$ for if not, $\bar{\eta}(x) < x \leqslant \bar{\eta}(x)$ which is a contradiction. Also, $x > y^{*n} \wedge y^{*n+1}$ for if not, $\bar{\eta}(z^{*n} - F_n^{-1}(0)) \leqslant \bar{\eta}(y^{*n}) < y^{*n+1}$. Then by Lemmas 3 and 4, $y^{*n} = y_n^* \geqslant y_{n+1}^* \geqslant y^{*n+1} > \bar{\eta}(y^{*n}) \geqslant y^{*n}$, a contradiction. Thus, we have $y^{*n} \wedge y^{*n+1} < x < y^{*n} \vee y^{*n+1}$. Now if $y^{*n} < y^{*n+1}$, then $\bar{\eta}(y^{*n}) < \bar{\eta}(x) < y^{*n+1}$, so by the previous argument we have a contradiction. If instead $y^{*n+1} < y^{*n}$, then $y^{*n} \leqslant \bar{\eta}(y^{*n}) < x < y^{*n}$, a contradiction. This completes the proof.

COROLLARY 2′. If either $y_n^* \leqslant y_{n+1}^*$ and $y^{*n} > y^{*n+1}$, or $y_n^* < y_{n+1}^*$ and $y^{*n} \geqslant y^{*n+1}$, then $\eta(z^{*n+1} - F_{n+1}^{-1}(0)) \geqslant y^{*n+2}$ and $\bar{\eta}(\bar{y}^{n+1}(x)) \geqslant \bar{y}^{n+2}(x)$ for all x.

PROOF: Assume $\eta(z^{*n+1} - F_{n+1}^{-1}(0)) < y^{*n+2}$. By Lemmas 3 and 4, $y^{*n+1} = y_{n+1}^* \geqslant y_n^* \geqslant y^{*n} \geqslant y^{*n+1}$, so equality occurs throughout. This contradicts $y^{*n} > y^{*n+1}$ in the first case and $y_n^* < y_{n+1}^*$ in the second. Suppose there is an x such that $\bar{\eta}(\bar{y}^{n+1}(x)) < \bar{y}^{n+2}(x)$. We must have $x < y^{*n+1} \vee y^{*n+2}$ for if not, $\bar{\eta}(x) < x \leqslant \bar{\eta}(x)$, a contradiction. Also, $x > y^{*n+1} \wedge y^{*n+2}$, for if not, $\bar{\eta}(y^{*n+1}) < y^{*n+2}$, a contradiction. Thus, $y^{*n+1} \wedge y^{*n+2} < x < y^{*n+1} \vee y^{*n+2}$. Now if $y^{*n+1} < y^{*n+2}$, then $\bar{\eta}(y^{*n+1}) \leqslant \bar{\eta}(x) < y^{*n+2}$, a contradiction. If instead $y^{*n+1} > y^{*n+2}$, then $y^{*n+1} \leqslant \bar{\eta}(y^{*n+1}) < x < y^{*n+1}$, a contradiction, *completing the proof.*

THEOREM 8.

PROOF: If $P_n \leqslant P_{n+1}$, then by Theorem 3, $z^*(P_n) \leqslant z^*(P_{n+1})$. Thus, if $y^*(P_n) \geqslant 0$ for all n and $\eta_+ \leqslant 1$, then $\eta(z^*(P_n) - F_n^{-1}(0)) \leqslant z^*(P_n)^+ \leqslant z^*(P_{n+1})^+ \leqslant y^*(P_{n+1})$. If $\eta_+ = \eta_- = 1$, then $\eta(z^*(P_n) - F_n^{-1}(0)) \leqslant z^*(P_n) \leqslant z^*(P_{n+1)} \leqslant y^*(P_{n+1})$. Similarly, if $P_n \geqslant P_{n+1}$, then by Theorem 3, $y^*(P_n) \leqslant y^*(P_{n+1})$, so $\eta(z^*(P_n) - F_n^{-1}(0)) \leqslant \eta(y^*(P_n)) \leqslant y^*(P_n) \leqslant y^*(P_{n+1})$ by Condition (IV) and η nondecreasing. Thus in all cases, $\eta(z^*(P_n) - F_n^{-1}(0)) \leqslant y^*(P_{n+1})$, whence by a result of Veinott [18], the myopic policy is optimal for $x \leqslant y^*(P_1)$. Thus $(y,\beta) = (y^*(P_n), \beta^*(P_n))$ minimizes the term in braces in (2) subject to $\beta \geqslant 0$. But $(y^*(P^n), \beta^*(P^n))$ is the least such minimum so $y^*(P^n) \leqslant y^*(P_n)$ and $\beta^*(P^n) \leqslant \beta^*(P_n)$. And by Lemma 3, $\beta^*(P^n) \geqslant \beta^*(P_n)$, so $\beta^*(P^n) = \beta^*(P_n)$. Also, $EC[\eta(y - \beta^*(P^n) - U); P^{n+1}]$ is constant for $y \leqslant y^*(P_n)$ so $(y,\beta) = (y^*(P^n), \beta^*(P^n))$ minimizes $H(y) + P_n(\beta) + S(y - \beta)$ subject to $\beta \geqslant 0$. But $(y^*(P_n), \beta^*(P_n))$ is the least such minimum, so $y^*(P_n) \leqslant y^*(P^n)$ and, hence, $(y^*(P^n), \beta^*(P^n), z^*(P^n)) = (y^*(P_n), \beta^*(P_n), z^*(P_n))$, completing the proof.

In Theorem 9, $\mathscr{X}^{*T}$ is respectively $(\mathscr{X}^*_{n_1}, \mathscr{X}^*_{n_1+1}, \ldots, \mathscr{X}^*_{n_2})$ where $T = [n_1, n_2]$. In Theorem 10, $\mathscr{S}^T$ is defined similarly.

THEOREM 9.

PROOF: By Theorem 2, $y^*(H_n)$ and $z^*(H_n)$ are unimodal in n on T such that the modes of H_n on $\mathscr{X}^{*T}$ are modes of $y^*(H_n)$ and $z^*(\iota i_n)$. Thus, from Theorems 6 and 7, $y^*(H^n)$ and $z^*(H^n)$ are unimodal in n on T such that $m', m'' \leqslant m$. If $m' < m''$, then $z^*(H^n) < z^*(H^{n+1})$ for $n \equiv m'' - 1 < m$, so by Condition (IV), Lemma 3, and Lemma 5, $\beta^*(H^n) = \beta^*(H_n)$. Thus, by Theorem A2 and Lemma 5, $\beta^*(H^n) = \beta^*(H_n) \leqslant \beta^*(H_{n+1}) \leqslant \beta^*(H^{n+1})$. But since $m' \leqslant n$, $y^*(H^n) \geqslant y^*(H^{n+1})$, so $z^*(H^n) \geqslant z^*(H^{n+1})$ which is a contradiction. Therefore, $m'' \leqslant m'$.

Now by Theorems 6 and 7, $(y^*(H^n), \beta^*(H^n), z^*(H^n)) = (y^*(H_n), \beta^*(H_n), z^*(H_n))$ for $n < m''$. Thus, by Theorem 2 and Lemma 4, $\beta^*(H^n) = \beta^*(H_n) \leqslant \beta^*(H_{n+1}) \leqslant \beta^*(H^{n+1})$ for $n < m''$. It remains only to show $\beta^*(H^n)$ is nondecreasing in n for $m'' \leqslant n \leqslant m$. To this end consider

$$Q(z; H_n) \equiv \min_{y-\beta=z,\beta\geqslant 0} \{H_n(y) + P(\beta)\}. \tag{15}$$

Since by Lemma 1, $H_n(y)$ is subadditive in (y, H_n) on $R \times \mathscr{X}'^*_n$, the least element $(y, \beta) = (\hat{y}(z; H_n), \hat{\beta}(z; H_n))$ minimizing the term inside the braces in (15) over the indicated sublattice is nonincreasing in H_n on $\mathscr{X}_n$ by Theorem A2. Replacing y by $\beta - (-z)$ in (15), one sees from the convexity of H_n and example A6 that the new term inside the braces is subadditive in $(\beta, -z)$ on the sublattice $\{\beta \geqslant 0\}$. Thus by Theorem A2, $\hat{\beta}(z; H_n)$ is nonincreasing in z. Because S and C^{n+1} in recursion (2) are functions of z alone, $\hat{\beta}(z^*(H^n); H_n) = \beta^*(H^n)$. Thus, for $m'' \leqslant n < m$, $z^*(H^n) \geqslant z^*(H^{n+1})$ and $H^n \geqslant H^{n+1}$, so $\beta^*(H^n) \leqslant \beta^*(H^{n+1})$, completing the proof.

THEOREM 10.

PROOF: By Theorem 4 and 7, $z^*(S^n)$ is unimodal in n on T with least mode not exceeding that of S_n on $\mathscr{S}^{*T}$. Now $H(y) + P(y - z)$ is subadditive in (y,z) by the convexity of P and example A6. Therefore, by Theorem A2, the least $y = \hat{y}(z)$ minimizing $H(y) + P(y - z)$ subject to $y \geqslant z$ is nondecreasing in z. Thus, $y^*(S^n) = \hat{y}(z^*(S^n))$ is nondecreasing in $z^*(S^n)$ so $y^*(S^n)$ is unimodal in n and each mode of $z^*(S^n)$ is a mode of $y^*(S^n)$.

Now $H(z + \beta) + P(\beta)$ is superadditive in (z, β) by the convexity of H and example A6. Therefore, by Theorem A2, the least $\beta = \hat{\beta}(z)$ minimizing $H(z + \beta) + P(\beta)$ subject to $\beta \geqslant 0$ is nonincreasing in z. Hence, $\beta^*(S^n) = \hat{\beta}(z^*(S^n))$ is nonincreasing in $z^*(S^n)$ so $-\beta^*(S^n)$ is unimodal in n and each mode of $z^*(S^n)$ is a mode of $-\beta^*(S^n)$, completing the proof.

MINIMIZING THE AVERAGE DEVIATION OF JOB COMPLETION TIMES ABOUT A COMMON DUE DATE

John J. Kanet

The University of Georgia
Athens, Georgia

ABSTRACT

This paper considers a single-machine scheduling problem in which penalties occur when a job is completed early or late. The objective is to minimize the total penalty subject to restrictive assumptions on the due dates and penalty functions for jobs. A procedure is presented for finding an optimal schedule.

INTRODUCTION

For the most part, the literature of scheduling has been confined to problems involving penalty functions which are nondecreasing in job completion times. Conway, Maxwell, and Miller [2, p.12] refer to such functions as regular performance criteria. There are, however, many applications in which nonregular criteria are appropriate. Applications of nonregular measures occur, for example, in file organization problems. As indicated by Merten and Muller [4], minimizing the variance of retrieval times for records in a file may be highly desirable, especially in on-line systems. In spite of the importance of nonregular performance measures, very little analytical work has been done in this area. This is primarily due to the difficulty of solving this type of problem. The problem of minimizing completion time variance has been studied by Merton and Muller [4], Schrage [5] and Eilon and Chowdhury [3]. Merton and Muller [4] have analyzed the relation between flowtime and waiting time variance. They do not show how to minimize such measures but do show that if some schedule S minimizes (maximizes) one of the measures, then the antithetical schedule S' minimizes (maximizes) the other measure. They also show that the minimum value of both measures is the same. Schrage [5] has examined scheduling for minimum completion time variance when there are up to 5 jobs to be scheduled. Eilon and Chowdhury [3] have extended Schrage's work by showing that for a schedule to have minimum completion time variance it must be V-shaped. Let job k be the job with the smallest processing time of all the jobs to be scheduled. A schedule is V-shaped if the jobs placed before job k are in descending order of processing time and the jobs placed after job k are in ascending order of processing time. Eilon and Chowdhury [3] then compare various heuristics for minimizing completion time variance.

Apart from the problem of minimizing completion time variation are those problems in which the nonregular measure is a function of job lateness. Consider, for example, a job shop which produces components for subsequent assembly into finished products. The due dates for components are based on the assembly schedule of the end products. If component orders are late then the assembly of a product may be delayed. The negative effect could be loss of assembly efficiency and customer good will. If a component order is completed early it must wait in storage until the production date of the product for which it is needed. The negative effect is an accumulation of component inventory.

Sidney has addressed the type of problem noted above and in [6] he presents an algorithm for minimizing the maximum penalty when penalties are incurred both when jobs are completed early or late. In this paper we continue the work of Sidney by presenting a simple $[0(n^2)]$ algorithm for minimizing total cost when costs increase linearly as a job's completion date deviates from its due date.

THE PROBLEM

Consider a single machine with n jobs immediately available for processing. Associated with each job i is its required processing time p_i. All jobs have the due date $d \geqslant MS$ where MS is the makespan of the job set; i.e.,

$$MS = \sum_{i=1}^{n} p_i.$$

The objective is to find a schedule S which minimizes

$$(1) \qquad Z(S) = \sum_{i=1}^{n} abs(C_i - d)/n,$$

where C_i represents the completion time for job i and abs denotes the absolute value function. Equation 1 is the mean absolute lateness (MAL) resulting from schedule S.

DETERMINING AN OPTIMAL SCHEDULE

If preemption is allowed, the problem is trivial. In that case an optimal solution with $MAL = 0$ can always be obtained by processing the jobs in any order during the time interval $(d - MS, d)$. Every job is interrupted when it has an arbitrarily small amount of processing time ϵ remaining. After the initial processing of all jobs, the remaining $n\in$ amount of work is completed. Since ϵ can be made arbitrarily small, the completion date of each job can be made to converge to d, that is, for each job i

$$\lim_{\epsilon \to 0} (abs[C_i - d]) = 0.$$

Thus in the limit, MAL = 0.

Suppose that preemption is not allowed. Then let B represent an ordered set of jobs to be scheduled without inserted idle time such that the last job in B is completed at time $t = d$. Let A represent an ordered set of jobs to be scheduled without inserted idle time such that the first job in A starts at time $t = d$. The following algorithm produces an optimal schedule S, defined by the permutation (B, A). Let U denote the set of unscheduled jobs. The symbol Φ denotes the empty set.

```
procedure SCHED:

    $B \leftarrow A \leftarrow \Phi$;
    while $(U \neq \Phi)$ do
        remove a job $k$ from $U$ such that $p_k = \max_i \{p_i\}$;
        insert job $k$ into the last position in $B$;
        if ($U \neq \Phi$) do
           remove a job $k$ from $U$ such that $p_k = \max_i \{p_i\}$;
           insert job $k$ into the first position in $A$;
        end
    end
    $S \leftarrow (B, A)$;
end SCHED
```

Figure 1 illustrates how the procedure works. In the sample problem there are five jobs with processing times given. The due date for each job is $t = 39$. The five Gantt charts illustrate how the algorithm progressively assigns jobs to the two sets B and A.

Problem data:

job identification :	J1	J2	J3	J4	J5
processing time :	7	12	5	4	10

due date = 39

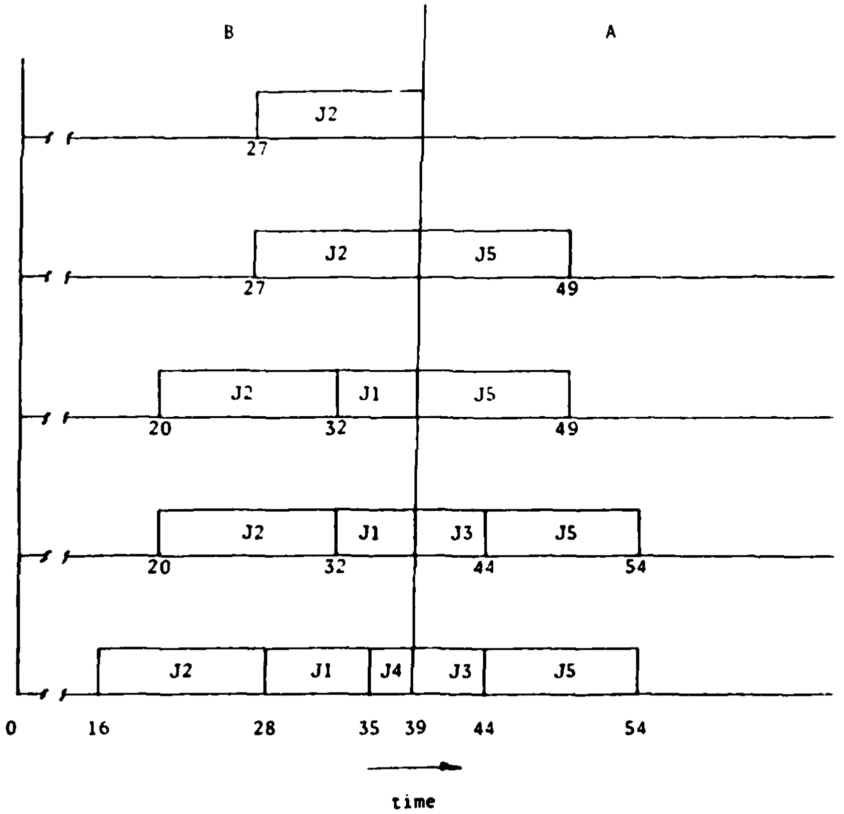


FIGURE 1. An example Problem and its solution by algorithm SCHED

PROOF OF THE ALGORITHM

To prove the correctness of procedure SCHED first observe that it is unnecessary to consider schedules that have idle time inserted between jobs in S. To show this is true, consider any schedule S with idle time inserted either before or after d. If the idle period occurs before (after) d, remove the idle time by moving the jobs before (after) the idle period forward (backward) in the schedule. The resulting schedule is an improvement over S since the completion dates of all jobs affected are moved closer to d. This process can be repeated until all inserted idle time is removed from the schedule.

A second observation is that schedules which have a job begin processing before d and end processing after d need not be considered. To show this let S be such a schedule. Let k be the job in S which is in process at time d (see Figure 2).

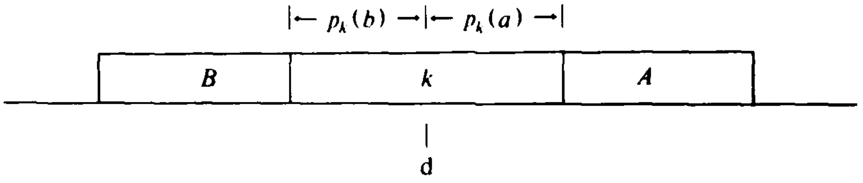


FIGURE 2.

Let $p_k(b)$ be the amount of processing time completed on job k before d and let $p_k(a)$ be the amount of processing time completed after d. Clearly, either $|B| \leqslant |A|$ or $|B| > |A|$ ($|B|$ denotes the cardinality of B). First suppose $|B| \leqslant |A|$. Then all jobs can be shifted to an earlier completion time such that $C_k = d$. The change in Z due to the shift is

$$|B|p_k(a) - |A|p_k(a) - p_k(a) < 0.$$

If $|B| > |A|$ all jobs can be shifted to a later completion time such that $C_k = d + p_k$. The change in Z due to this shift is

$$|A|p_k(b) - |B|p_k(b) + p_k(b) \leqslant 0.$$

In either event the objective function cannot increase by such a shift.

Because of the above two observations and because preemption is not allowed, it is sufficient to confine the search for an optimal schedule to the set of permutation schedules formed by the concatenation of B to A. By specifying such a permutation, it is understood that no idle time shall appear between jobs and that the last job in B will be completed at time $t = d$.

Observe that the following properties characterize the schedule S produced by SCHED:

1. The jobs in B are sequenced by longest processing time first (LPT),
   the jobs in A are sequenced by shortest processing time first (SPT);
2. If n is even $|B| = |A|$,
   if n is odd $|B| = |A| + 1$;
3. There is a one-to-one mapping of the jobs in A into the jobs in B such that

$$k \in A \rightarrow j \in B \Rightarrow p_k \leqslant p_j.$$

To prove that SCHED yields optimal solutions we must show that the above properties are both necessary and sufficient for a schedule to be optimal. We begin with the condition of necessity by assuming that some schedule S^* with these properties is not optimal. Then it must be that another schedule $S \neq S^*$ is optimal. Note that because there may be jobs with identical processing times there may be many schedules satisfying properties 1, 2, and 3 that have the same value for Z. S is none of these sequences. Then for $S \neq S^*$ it must be that at least one of the three properties is not met for S.

First consider property 1. Suppose for schedule S that B is not in LPT sequence or that A is not in SPT sequence. Assume the former condition is true. The proof of the latter is similar. If B is not in LPT sequence then there exists two adjacent jobs j, k such that $p_j < p_k$ (See Figure 3).

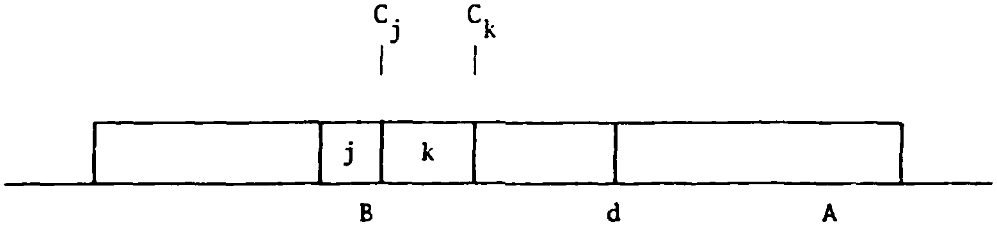


FIGURE 3.

Consider interchanging jobs j and k to form S'. Clearly, such an exchange does not affect the penalty for any job other than j or k. The change in Z produced by such an exchange is

$$Z' - Z = (d - C_j') - (d - C_j) + (d - C_k') - (d - C_k). \tag{2}$$

The first two terms of Equation 2 represent the change in penalty for job j; the second two terms are the change in penalty for job k. Substituting

$C_j' = C_k$, $C_k' = C_k - p_j$, and $C_j = C_k - p_k$ yields

$$Z' - Z = p_j - p_k < 0.$$

Thus, $Z(S') < Z(S)$, contrary to the assumption that S is optimal.

Next, assume that $S \neq S^*$ because property 2 is not met for S. For this to be so, it must be that either $|B| < |A|$ or $|B| > |A| + 1$.

CASE 1: $|B| < |A|$

Let k be the smallest job in A. By property 1, k must be the first job in A. Form schedule S' by removing k from A and making it the last job in B. The resulting change in Z is

$$Z' - Z = p_k|B| - p_k|A|. \tag{3}$$

The first term in Equation 3 is the increase in penalty for the jobs originally in B. The second term is the decrease in penalty for the jobs originally in A. Without loss of generality assume $|B| = |A| - 1$. Then

$$Z' - Z = -p_k < 0,$$

which is contrary to the assumption that S is optimal.

CASE 2: $|B| > |A + 1$

Let k be the smallest job in B. By property 1, k must be the last job in B. Form schedule S' by removing k from B and making it the first job in A. The resulting change to Z is

$$Z' - Z = p_k|A| - p_k(|B| - 1) + p_k. \tag{4}$$

The first term in Equation 4 is the increase in penalty for the jobs in A. The second term is the decrease in penalty for all jobs in B other than k. The third term represents that net change to job k. Since $|B| > |A| + 1$ it follows that

$$Z' - Z \leqslant 0.$$

Clearly, if $Z' - Z \leqslant 0$, then S cannot be optimal. If $Z' - Z = 0$, then it must be that $|B| = |A| + 2$, allowing S to be transformed into an equivalent schedule satisfying property 2.

Finally, consider property 3. Assume that $S \neq S^*$ because no mapping satisfying property 3 exists for S. Note that if S is optimal, properties 1 and 2 must hold. Attempt to construct a mapping satisfying property 3 by the following procedure. Start with the largest (last) job in A and match it to the largest (first) job in B. Continue with the second largest job and so on until the next job k in A to be matched to j in B cannot be done because $p_k > p_j$. The attempt to map jobs must fail in this fashion, otherwise there would be a mapping contrary to assumption. Let BJ be the jobs that precede j and let AK be the jobs that follow k in S (see Figure 4). Clearly $|BJ| = |AK|$. Now form S' by interchanging jobs j and k. This affects the penalty of j and k and the jobs in BJ and AK, but no other jobs.

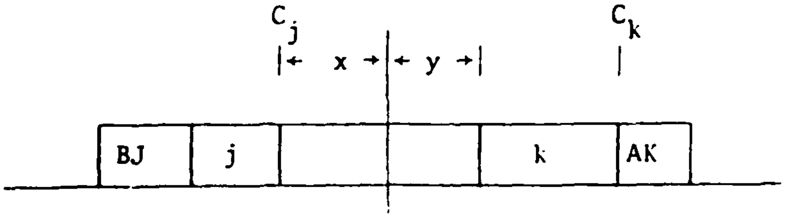


FIGURE 4.

Let $x = d - C_j$ and $y = C_k - p_k - d$. The change in penalty caused by forming S' is

$$Z' - Z = |BJ|(p_k - p_j) + |AK|(p_j - p_k) + (x - y - p_k) + (y + p_j - x). \tag{5}$$

The first term of Equation 5 is the change in penalty for the jobs in *BJ*. The second term is the change for jobs in *AK*. The third term is the change to job *k*; and the fourth term is the change to job *j*. Equation 5 simplifies to

$$Z' - Z = p_j - p_k < 0,$$

which contradicts the assumption that S is optimal. This completes the arguments for the necessity of properties 1-3. To show that the three properties are sufficient to define an optimal schedule assume that S' exists which satisfies the three properties and

$$(6) \qquad Z(S') > Z(S)$$

where S is an optimal schedule. If S is optimal then it too must satisfy the three properties. Now

$$Z(S') > Z(S) \Rightarrow S' \neq S$$

but $S' = S$ since the three properties are sufficient to define a permutation of the jobs. This contradicts (6) and completes the proof.

SIMILARITY TO THE 2-MACHINE $\bar{F}$ PROBLEM

At first glance one might mistakenly believe that the problem defined here can be reduced to the 2-machine $\bar{F}$ problem discussed by Baker [1, p. 118] among others (e.g., Conway, Maxwell, and Miller [2, p. 74]). By referring again to procedure SCHED and Figure 1 we can see how such a reduction might be construed. We simply let B and A represent the two machines and assume that the jobs assigned to these two machines are sequenced in SPT order for each machine. The example job set shown in Figure 1 would then be sequenced as indicated in Figure 5.

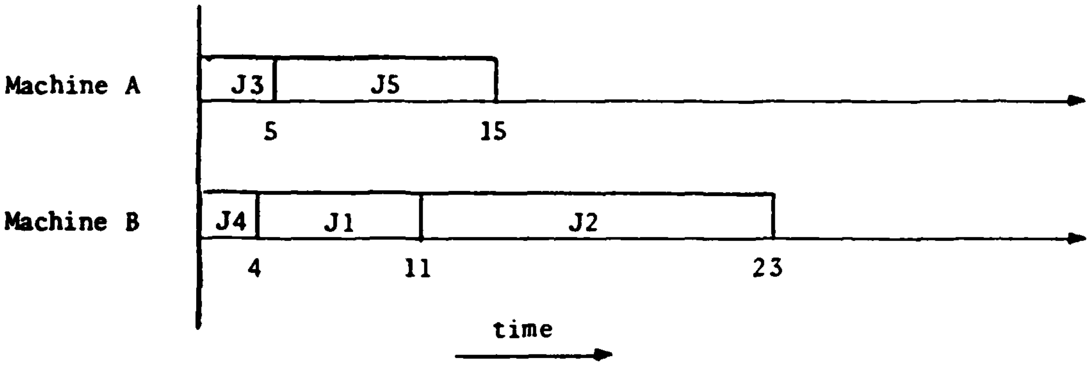
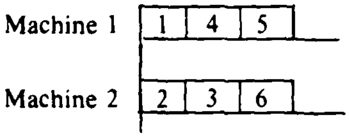


FIGURE 5.

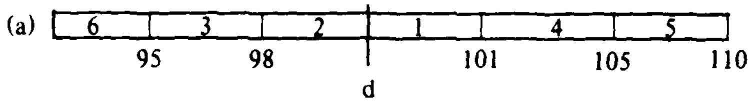
If the MAL problem presented here were equivalent to the 2-machine $\bar{F}$ problem then any optimal solution to one would also be optimal for the other. This is not so and to see why consider the example problem appearing in Baker [1, pp. 118-119] and given below:

job	j	1	2	3	4	5	6
	p_j	1	2	3	4	5	6

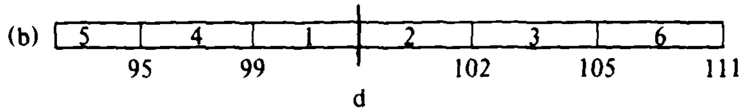
As Baker points out, one optimal solution to the $\bar{F}$ problem is (not to scale):



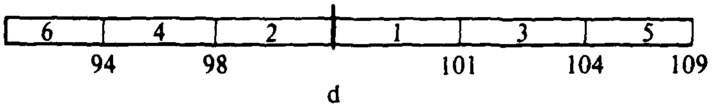
Let us assume a due date of $d = 100$. One could then "interpret" the above solution as either:



or



Case (a) yields MAL of 23/6. Case (b) yields MAL = 24/6. Procedure SCHED yields the optimal schedule:



which has an MAL of 22/6. Clearly then an optimal solution to the $\bar{F}$ problem is not always an optimal solution to the MAL problem. Therefore, the two problems are different. The reason for this lies in the fact that property 3 (given above) is necessary for an optimal solution to the MAL problem but not necessary for the $\bar{F}$ problem.

SUMMARY

What we have addressed here are problems in which penalties for jobs are incurred both when they are completed early or late. The critical condition that $d \geqslant$ MS is sufficient to ensure that procedure SCHED does not require jobs to be processed prior to $t = 0$. It has been shown that under this assumption the optimal schedule must be *V*-shaped. Whereas procedure SCHED provides optimal solutions under the noted conditions, future research must be directed toward problems in which d is not restricted and to the case of multiple due dates.

REFERENCES

[1] Baker, K., *Introduction to Sequencing and Scheduling* (John Wiley & Sons, New York, 1974).

[2] Conway, R.W., W.L. Maxwell, and L.W. Miller, *Theory of Scheduling* (Addison-Wesley, Reading, MA, 1967).

[3] Eilon, S. and I.G. Chowdhury, "Minimising Waiting Time Variance in the Single Machine Problem," Management Science, *23*, 6, 567-575 (1977).

[4] Merten, A.G. and M.E. Muller, "Variance Minimization in Single Machine Sequencing Problems," Management Science, *18*, 9, 518-528 (1972).

[5] Schrage, L., "Minimizing the Time-in-System Variance for a Finite Jobset," Management Science, *21*, 5, 540-543 (1975).

[6] Sidney, J.B., "Single-Machine Scheduling with Earliness and Tardiness Penalties," Operations Research, *25*, 62-69 (1977).

A MATHEMATICAL MODEL FOR GENERATING THE AREA OF A DRAINAGE BASIN

R. T. Robinson

Engineer Studies Center
Corps of Engineers
Washington, DC

ABSTRACT

This paper presents a mathematical model that yields the area drained by a naturally-occurring network of streams. The model is based on empirically derived relationships in the field of quantitative geomorphology and an assumption concerning the probabilistic nature of stream system formation. A wide range of model solutions is indicated, and the model is validated by comparing the results to statistics from actual stream systems.

INTRODUCTION

The areas drained by stream systems are necessary factors in the solution of many problems in engineering design and military planning. Securing such areas using topographic maps, aerial photographs, or actual measurement is often fraught with difficulties related to such factors as time, money, and access. This paper presents a model that was developed to satisfy such data requirements analytically through the quantification of a few simple parameters in the geographic areas of interest. The model is based on empirically derived relationships in the field of quantitative geomorphology and assumptions concerning the probabilistic nature of stream system formation. Model solutions are indicated for stream systems of various sizes and model validity is examined using statistics from actual stream systems.

BACKGROUND

The study of the formation of drainage patterns from a stochastic viewpoint had an early beginning. In 1802, John Playfair, Professor of Natural Philosophy, University of Edinburgh, wrote *Illustrations of the Huttonian Theory of the Earth.* The following passage has been called Playfair's Law:

> "Every river appears to consist of a main trunk, fed from a variety of branches, each running in a valley proportioned to its size, and all of them together forming a system of valleys, communicating with one another, and having such a nice adjustment of their declivities that none of them join the principal valley either on too high or too low a level; a circumstance which would be infinitely improbably if each of these valleys were not the work of the stream which flows in it."

Playfair obviously recognized the systematic aspects of a natural drainage system; however, little was done to quantify these aspects prior to the work of R. E. Horton in 1945 [3]. Horton's bifurcation and stream length ratios are particularly important to this model and to most of the work that has been accomplished in the field since Horton's publication.

In recent years, probability theory has been used in an attempt to explain and quantify various aspects of naturally occurring drainage networks. Prominent among the works in this area have been those by Hack [2], Leopold and Langbein [4] and Schenck [8]. The general theory basic to these works is that the formation of drainage networks can be explained by the laws of probability and that, free of constraints, streams tend to occur in a random fashion. Leopold and Langbein viewed the laws of Horton as representing the most probable state in a stochastically formed network and used a random walk model to show that the laws of Horton did, in fact, hold. Schenck used a similar model to substantiate the empirical work of Gray [1].

The first known attempt at constructing a predictive model was made by Mayer [5]. The Mayer model was an attempt to use the work of the several previous authors in the field to construct drainage basin relationships that, with quantification of a few parameters, could be used to predict the numbers, orders, and hydrologic characteristics of streams within a naturally occurring drainage network. The model described in this paper has some of the characteristics of the Mayer model.

BASIC CONCEPTS

Basic to the model is the concept of a drainage basin. Loosely defined, a drainage basin is the total area drained by a stream and its tributaries. Every drainage basin is circumscribed by a drainage divide that separates the precipitation that falls into that basin from the precipitation that falls into contiguous basins. Drainage basin areas of the same order are nonoverlapping; and the total area drained by all basins exhausts the surface of the earth's landmass. These concepts have long been used in the study of stream morphology and are basic to the relationships described in this paper.

Also basic to the model is the Horton method of stream ordering. This method traces each stream to its drainage divide and considers that each stream extends from its mouth to primary headwater. The method is illustrated in Figure 1.

BASIC RELATIONSHIPS

The model presented in this paper draws heavily on four empirically established relationships—two derived by Horton [3], one derived by Schumm [9], and one derived from works by Hack [2]. Although the relationships are least-squares approximations, they represent data collected over wide geographic areas, and the degrees of fit were generally quite good.

The Horton relationships may be stated as follows:

(1) $l_i = l_1 \alpha^{i-1}$,

(2) $n_i = \beta^{N-i}$,

where

l_i = mean length of an ith-order stream,

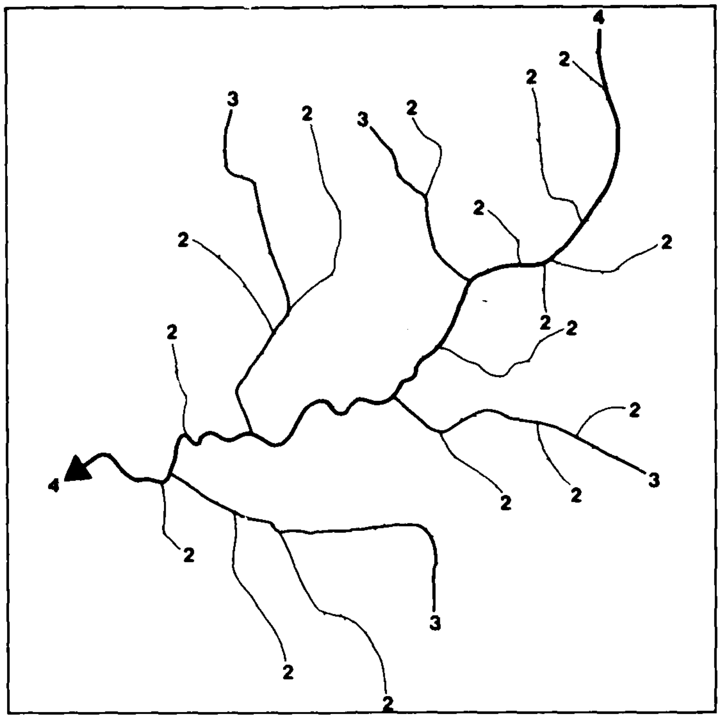


FIGURE 1. Horton definition of stream order (redrawn from Shreve [10]). Order indicated by number near upstream end of respective streams. Unnumbered streams are first order.

n_i = number of ith-order streams in an Nth-order drainage basin,

α = stream length ratio, a constant factor derived from empirical evidence by which the mean length of an ith-order stream exceeds the mean length of an $(i-1)$st-order stream,

and

β = bifurcation ratio, a constant factor derived from empirical evidence by which the number of $(i-1)$st-order streams exceeds the number of ith-order streams in an Nth-order drainage basin.

In examining collections of data, Horton and others found that the values observed for α and β varied somewhat from basin to basin, depending on such factors as gradient, soil composition, rainfall, etc., but that the values recorded in mature basins tended to cluster around 2.68 and 4, respectively. Consequently, these values are referred to throughout this paper as the equilibrium values for these basic parameters.

The Schumm relationship may be stated as follows:

(3) $$A_{N-j} = A_N \lambda^j,$$

where

A_i = area drained by all ith-order basins in an Nth-order parent basin,

and

λ = basin area ratio, a constant factor by which the total area drained by all basins of ith-order exceeds the total area drained by all basins of $(i - 1)$st-order in an Nth-order drainage basin.

The Hack relationship may be stated as follows:

(4) $$f_i = f_1\alpha^{i-1},$$

where

f_i = area drained by overland flow occurring directly into an ith-order stream,

and α is the stream length ratio defined by Horton. Note that the area drained by overland flow is that area for which the sheet flow is insufficient in length to sustain a first-order stream; e.g., the area immediately adjacent to the banks of a stream.

Using the Hack and Schumm relationships together, we can also form the useful relationship:

(5) $$F_N = (\lambda\alpha/\beta)^{N-1}A_N,$$

where

F_N = total area drained by overland flow occurring directly into the Nth-order stream.

THE JUMPING PROCESS

A phenomenon of particular importance in describing the behavior of streams is the process of "jumping," a phenomenon often ignored in constructing drainage system models. Specifically, a stream of order i is said to jump if at its mouth it intersects a stream of order $i + 2$ or greater; i.e., the stream does not bifurcate with the next higher order stream but jumps at least one order in the hierarchy. To describe this process analytically, the following assumption is made:

ASSUMPTION: Within an Nth order parent basin, streams of order i bifurcate with streams of order $i + 1$ and greater in direct proportion to the relative total stream length of the recipient streams. Based on this assumption, the proportion of streams in an Nth-order parent basin that jump to the order N stream may be derived as follows:

First, the number of ith-order streams that bifurcate with $(i + 1)$st-order streams may be written as

$$n_{i,i+1} = n_i n_{i+1} l_{i+1} \Big/ \sum_{j=i+1}^{N} n_j l_j.$$

Dividing this quantity by n_i and taking its complement yields the proportion of ith-order streams that bifurcates with all orders greater than $i + 1$ (i.e., the proportion that jumps). This proportion, denoted P_i, may be written as

$$P_i = 1 - n_{i+1} l_{i+1} \Big/ \sum_{j=i+1}^{N} n_j l_j,$$

which, using equations (1) and (2), sums to

$$P_i = \alpha(\beta^{N-i-1} - \alpha^{N-i-1}) / (\beta^{N-i} - \alpha^{N-i}).$$

Next, of the ith-order streams that jump, the proportion that jumps specifically to the parent stream of order $N \geqslant i + 2$, denoted $P_{i,N}$, may be written as

$$P_{i,N} = n_N l_N / \sum_{j=i+2}^{N} n_j l_j,$$

which, using equations (1) and (2), sums to

$$P_{i,N} = \alpha^{N-i-2}(\beta - \alpha)/(\beta^{N-i-1} - \alpha^{N-i-1}).$$

Finally, the proportion of all ith-order streams that jumps to the order N stream may be expressed as

(6) $$P_i P_{i,N} = \alpha^{N-i-1}(\beta - \alpha)/(\beta^{N-i} - \alpha^{N-i}).$$

Solutions of this equation for parent basins of orders 3 through 9 and equilibrium α, β values are set forth in Table 1.

TABLE 1 — *Proportion of Streams in a Drainage Basin That Jump to Parent Stream*

Order of Jumping Stream (i)	Order of Parent Stream (N)						
	3	4	5	6	7	8	9
1	.401	.212	.124	.077	.049	.032	.021
2	0	.401	.212	.124	.077	.049	.032
2	0	0	.401	.212	.124	.077	.049
4	0	0	0	.401	.212	.124	.077
5	0	0	0	0	.401	.212	.124
6	0	0	0	0	0	.401	.212
7	0	0	0	0	0	0	.401

NOTE: Values were computed using $\alpha = 2.68$ and $\beta = 4$, the equilibrium values.

THE FUNDAMENTAL RELATIONSHIP

In examining collections of field data, Hack [2] found that the area of a basin could be related to its length and shape by the power function

$$l_i = k a_i^h,$$

where l_i represents the mean length of ith-order streams, a_i the mean area of ith-order basins, h a parameter indicating the shape of the mean basin, and k a constant factor by which the length of the ith-order stream exceeds the length of the ith-order basin. Assuming that h and k are constant in relatively homogeneous basins of adjacent order, we can write

$$\begin{aligned}\alpha &= l_i/l_{i-1},\\ &= [a_i/a_{i-1}]^h,\end{aligned}$$

which, multiplying both sides by $1/\beta$, reduces to the fundamental relationship

$$\beta = \lambda\alpha^{1/h}. \tag{7}$$

It can be shown that h is logically bounded by $0.5 \leqslant h \leqslant 1.0$. However, Hack [2], Mayer [5], and others found that values not exceeding 0.6 were typical for h in relatively mature basins.

AN ADMISSABLE REGION FOR MODEL PARAMETERS

The area of a parent basin, denoted A_N, may be written as the sum of three components: the area drained by included basins of order $N-1$ that bifurcate directly with the order N stream; the area drained by the basins of orders $N-2$ and lower that jump directly to the order N stream; and the area drained by overland flow occurring directly into the order N stream. Symbolically, this relationship may be written as

$$A_N = A_{N-1} + P_{N-2}P_{N-2,N}A_{N-2} + \cdots + P_1P_{1,N}A_1 + F_N,$$

which, using equations (3), (5), and (6), becomes

$$A_N = \lambda A_N + A_N\left(\frac{\beta-\alpha}{\alpha}\right)\sum_{j=2}^{N-1}\left[\frac{(\lambda/\beta)^j}{\alpha^{-j}-\beta^{-j}}\right] + A_N(\lambda\alpha/\beta)^{N-1},$$

and further reduces to the basic parametric relationship

$$\left(\frac{\beta-\alpha}{\alpha}\right)\sum_{j=2}^{N-1}\left[\frac{(\lambda/\beta)^j}{\alpha^{-j}-\beta^{-j}}\right] + \lambda + (\lambda\alpha/\beta)^{N-1} - 1 = 0. \tag{8}$$

This equation defines the equilibrium behavior of λ, the basin area ratio, within a relatively homogeneous drainage basin. The bounds on λ are determined by the bounds on P_i, the proportion of streams that jump. Hence, λ assumes its minimum value when P_i is maximum, i.e., $P_i = 1$, all $i < N$, and λ assumes its maximum value when $P_i = 0$, all $i < N$. With $P_i = 0$, all $i < N$, we obtain

$$\lambda + (\lambda\alpha/\beta)^{N-1} - 1 = 0, \tag{9}$$

which defines the upper bound on the admissable region for the α, β pairs. With $P_i = 1$, all $i < N$, we obtain

$$\beta\left(\frac{\beta-\alpha}{\alpha}\right)\sum_{j=2}^{N-1}\left[\frac{(\lambda/\beta)^j}{\alpha^{1-j}-\beta^{1-j}}\right] + \lambda + (\lambda\alpha/\beta)^{N-1} - 1 = 0, \tag{10}$$

which defines the lower bound on the admissable region for the α, β pairs. Using equation (7) to define the interdependence of the parameters, we can further write equations (8), (9) and (10) as follows:

$$\left(\frac{\beta-\alpha}{\alpha}\right)\sum_{j=2}^{N-1}\left[\frac{\alpha^{-j/h}}{\alpha^{-j}-\beta^{-j}}\right] + \beta\alpha^{-1/h} + \alpha^{(1/h)(h-1)(N-1)} - 1 = 0, \tag{11}$$

$$\beta\alpha^{-1/h} + \alpha^{(1/h)(h-1)(N-1)} - 1 = 0. \tag{12}$$

$$\beta\left(\frac{\beta-\alpha}{\alpha}\right)\sum_{j=2}^{N-1}\left[\frac{\alpha^{-j/h}}{\alpha^{1-j}-\beta^{1-j}}\right]+\beta\alpha^{-1/h}+\alpha^{(1/h)(h-1)(N-1)}-1=0. \tag{13}$$

Solutions of equations (11), (12) and (13) yield the admissable region for the α, β pairs for each parent basin and specified value for h, the basin shape parameter. The equations may be solved using such techniques as the Newton-Raphson method of successive approximation [7]. Solutions for $N = 9$ and $h = 0.55$ are illustrated in Figure 2.

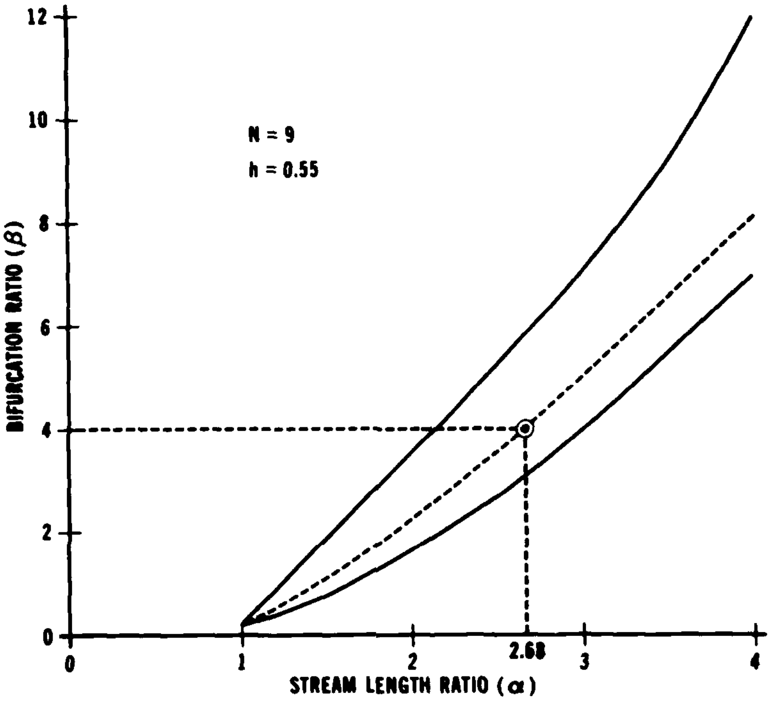


FIGURE 2. Admissable region for α, β pairs (the middle curve is the locus of equilibrium values)

BASIN AREA MODEL

As illustrated in the previous section, the area drained by a parent basin may be partitioned into three components. The recursive nature of this relationship is developed in Table 2 for parent basins of successively higher order. The development in Table 2 suggests that the area of an Nth-order, relatively homogeneous basin may be expressed in recursive form as follows:

$$A_N = \sum_{i=1}^{N}\beta^{N-i}f_i + \sum_{i=1}^{N-2}\sum_{j=i+2}^{N}\beta^{N-j}J_{i,j}. \tag{14}$$

Using equation (4), the Hack relationship for overland flow areas, the first term of equation (14) may be summed mathematically and expressed as:

TABLE 2 — *Recursive Relationship of Included Basin Areas*

Parent Basin Order (l)	Area of Parent Basin by Component			
	Orthodox Basins (A_{l-1})	Jumping Basins ($J_{k,l}$)	Overland Flow (f_l)	Total Area Drained (A_l)
1	0	0	f_1	f_1
2	βf_1	0	f_2	$\beta f_1 + f_2$
3	$\beta^2 f_1 + \beta f_2$	$J_{1,3}$	f_3	$\beta^2 f_1 + \beta f_2 + f_3 + J_{1,3}$
4	$\beta^3 f_1 + \beta^2 f_2 + \beta f_3 + \beta J_{1,3}$	$J_{1,4} + J_{2,4}$	f_4	$\beta^3 f_1 + \beta^2 f_2 + \beta f_3 + f_4 + \beta J_{1,3} + J_{1,4} + J_{2,4}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
N	$\sum_{i=1}^{N-1} \beta^{N-i} f_i + \sum_{i=1}^{N-3} \sum_{j=i+2}^{N-1} \beta^{N-j} J_{i,j}$	$\sum_{i=1}^{N-2} J_{i,N}$	f_N	$\sum_{i=1}^{N} \beta^{N-i} f_i + \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \beta^{N-j} J_{i,j}$

$$(15) \qquad \sum_{i=1}^{N} \beta^{N-i} f_i = f_1(\beta^N - \alpha^N)/(\beta - \alpha).$$

It may be observed that equation (15) is the fundamental Hack equation for an Nth-order basin area [2]. Hack did not consider the process of jumping in his development which made his area equation deficient by the jumping component. The second term of equation (14) may be expanded by considering the following development.

From the jumping process described in a previous section, the total area drained by a stream of order $k \geqslant 1$ that jumps to a parent stream of order $l \geqslant k + 2$ may be written as

$$J_{k,l} = n_k A_k P_k P_{k,l}.$$

That is, in any parent basin of order l there are n streams of order k, each of which has area A_k and a proportion equal to $P_k P_{k,l}$ that bifurcates directly with (jumps to) the order l stream. Therefore, $J_{k,l}$ is the total area drained by those kth-order streams that bifurcate directly with the parent (order l) stream. Using equation (2) for n_k and equations (14) and (15) for A_k, this relationship may be written in recursive form as

$$(16) \qquad J_{k,l} = P_k P_{k,l} \beta^{l-k} \left[f_1(\beta^k - \alpha^k)/(\beta - \alpha) + \sum_{\substack{i=1 \\ k>2}}^{k-2} \sum_{j=i+2}^{k} \beta^{k-j} J_{i,j} \right].$$

where $k \geqslant 1$ and $l \geqslant k + 2$.

Thus, the total area drained by an Nth-order basin may be written as

$$(17) \qquad A_N = f_1(\beta^N - \alpha^N)/(\beta - \alpha) + \sum_{\substack{k=1 \\ N>2}}^{N-2} \sum_{l=k+2}^{N} \beta^{N-l} J_{k,l}.$$

where $J_{k,l}$ is given by equation (16). Solutions of equation (17) for a range of α, β values are set forth in Table 3. Table 3 also reflects the implied values for λ, the basin area ratio, for adjacent equilibrium basins.

TABLE 3 — *Areas Drained by Parent Basins for a Range of Parameter Values*

Parent Stream Order (N)	α-10%	α+10%	Equilibrium Values	β+10%	β-10%	Basin Area Ratio for Equilibrium Values (λ)
	$\alpha=2.41$ $\beta=4$	$\alpha=2.95$ $\beta=4$	$\alpha=2.68$ $\beta=4$	$\alpha=2.68$ $\beta=4.4$	$\alpha=2.68$ $\beta=3.6$	
1	1.0	1.0	1.0	1.0	1.0	—
2	6.4	7.0	6.7	7.1	6.3	.60
3	37.5	43.3	40.3	45.7	35.3	.67
4	214.2	261.3	237.0	288.0	192.4	.68
5	1217.4	1559.3	1380.7	1804.9	1035.7	.69
6	7048.6	9459.5	8189.7	11512.1	5673.6	.67
7	42266.4	59456.1	50340.5	75911 7	32271.8	.65
8	263222.9	388819.2	321759.2	518871.1	191464.6	.63
9	1696276.2	2635843.5	2130301.0	3663590.5	1180099.3	.60

NOTE: Area values are in terms of f_1, the mean area drained by 1st-order streams, expressed in square miles.

The α, β values reflected in Table 3 were chosen to examine the sensitivity of the area model to these basic parameters. The values all lie within the admissable region depicted in Figure 2. The results show that the model is very sensitive to both α and β, but is most sensitive to β. For example, for $N = 9$, a 10 percent increase in β causes a 72 percent increase in basin area, whereas a 10 percent increase in α causes only a 24 percent increase in basin area. Sensitivity to both parameters also tends to increase as N increases. The sensitivity results suggest the importance of quantifying the model parameters in the area of interest, especially if there is reason to suspect the presence of nonequilibrium behavior.

The values for the basin area ratio, λ, in Table 3 were computed using equation (3). This equation may be written as $\lambda A_N = A_{N-1}$, where A_{N-1} is defined to be the total area drained by all $(N - 1)$st-order streams in an Nth-order parent basin. However, the areas given in Table 3 are for a single basin of each order. Therefore, in terms of the values in Table 3 we can write $\lambda A_N = \beta A_{N-1}$, which implies that λ differs from unity by the sum of two areas: the area drained by overland flow into the Nth-order steam and the area drained by those streams of orders $N - 2$ and lower that jump to the Nth-order stream. Thus, λ can be interpreted as the percent of the area drained by an Nth-order stream that is due to included basins of $(N - 1)$st-order. The balance $(1 - \lambda)$ is the proportion that is drained by overland flow into the Nth-order stream and the streams of order $N - 2$ and lower that jump directly to the Nth-order stream and are not included in $(N - 1)$st-order basins.

MODEL VALIDATION

The acid test of a model is the extent to which the model output correlates with observed data. In this case, comparison data are available for several of the major stream systems of the world. (See, for example, Horton [3] and Morisawa [6]). Table 4 lists relevant data for four of the world's major river systems and one smaller tributary of the Delaware River.

TABLE 4 — *Comparison of Model Results with Actual River Systems*

River System	River Length (Actual) (miles)	Area Drained (Actual) (sq miles)	Computed Order* (N)	Area Drained (Model) (sq miles)	Implied Mean Area Drained by First Order Streams* (sq miles) (f_1)
Delaware, East Branch	57	785	5	1,381 f_1	.57
Tennessee	900	40,600	7	50,341 f_1	.81
Arkansas	1450	160,500	8	321,759 f_1	.50
Mississippi	3892	1,243,700	9	2,130,301 f_1	.58
Nile	4160	1,150,000	9	2,130,301 f_1	.54

*See text for discussion.

The orders of the parent streams in Table 4 were computed by setting $i = N$ in equation (1) and writing the equation as follows:

$$N = \log(l_N/l_1)/\log\alpha + 1.$$

The values for N were then computed using $\alpha = 2.68$, its equilibrium value, and $l_1 = 1$. While l_1 tends to vary between about 0.5 and 1.5 miles depending on how well an area drains, the values tend to cluster about unity in mature basins. (See, for example, Hack [2] and Horton [3].) It may also be observed that both the areas drained and the lengths of the parent streams tend to vary somewhat within basin order since the basin orders are stated as integers and tend to increase in proportion to the logarithm of the areas drained.

A measure of the consistency with which the model estimates the actual areas is given in Table 4 by the implied mean values for the areas drained by first-order streams, f_1. Examinations of field data suggest that this value varies in nature between about 0.1 and 1.0 square miles, but that values of 0.5 to 0.8 are typical for mature basins. (See, for example, Hack [2], Horton [3], and Morisawa [6].) Thus, the model, solved for equilibrium parameter values, appears to produce very good approximations of actual basin area size. Better estimates could be expected if the parameters were quantified in the specific areas of interest; however, the effort required to obtain such estimates is often substantial.

A final measure that tends to validate the model is given by the λ values in Table 3. This can be seen by examining the relationship implied by the model results between β and the basin shape parameter, h. Since the area model, equation (17), is not a function of h, the λ values in Table 3 were generated independent of basin shape and thus could be considered to represent the equilibrium condition with respect to h. Equation (7), which relates λ to h, can be written as:

$$h = -\log\alpha/\log(\lambda/\beta).$$

By substituting the high and low λ values from Table 3 into this equation and using equilibrium α and β values, we obtain an implied range for h of 0.52 to 0.56 which correlates closely with values observed by Hack [2], Mayer [5], and others for the basin shape parameter; i.e., a basin that is slightly elongated in the downstream direction. It also correlates closely with the results plotted in Figure 2 which show that the equlibrium solution is attained when $\alpha = 2.68$, $\beta = 4$ and $h = 0.55$, and tends to validate the assumption made in deriving equation (7) that h is relatively constant in basins of adjacent order. Therefore, the model exhibits the desired balance and, since the results correspond closely to observed data, also tends to validate the underlying relationships empirically derived by Horton, Hack, and Schumm.

SUMMARY

The model described in this paper proved quite useful in estimating stream crossing requirements for military operations. Together with a companion model, the area equation was used to estimate the numbers, sizes, and interfluvial distances of streams in areas of interest. Although the model behaves quite well using equilibrium values for the model parameters, model behavior is enhanced by quantifying the parameters in the specific areas of interest. Recommendations to overprint these data on standard topographic maps were an outgrowth of this modeling effort.

REFERENCES

[1] Gray, D.M., "Interrelationships of Watershed Characteristics," Journal of Geophysical Research, *66*, 4 (1961).

[2] Hack, J.T., "Studies of Longitudinal Stream Profiles in Virginia and Maryland," Professional Paper 294-B, US Geological Survey, Washington, DC (1957).

[3] Horton, R.E., "Erosional Development of Streams and Their Drainage Basins: Hydrophysical Approach to Quantitative Morphology," Bulletin of the Geological Society of America, *56*, 275-370 (1945).

[4] Leopold, L.B. and W.B. Langbein, "The Concept of Entropy in Landscape Evolution," Professional Paper 500-A, US Geological Survey, Washington, DC (1962).

[5] Mayer, H.E., "Natural Drainage Systems," Vol. 1, Technical Operations, Inc., Combat Operations Research Group Memorandum CORG-M-363, CORG Project No. A7740, DA Contract No. DAAG-05-67-C-0547 (November 1968).

[6] Morisawa, M.E., "Relation of Quantitative Geomorphology to Stream Flow in Representative Watersheds of the Appalachian Plateau Province," Project No. 389-042, TR 20, Columbia University, New York (1959).

[7] Saaty, T.L. and J. Bram, *Nonlinear Mathematics* (McGraw-Hill Book Company, New York 1964).

[8] Schenck, H., Jr., "Simulation of the Evolution of Drainage Basin Networks with a Digital Computer," Journal of Geophysical Research, *68*, 20 (1963).

[9] Schumm, S.A., "Evolution of Drainage Systems and Slopes in Badlands at Perth Amboy, New Jersey," Office of Naval Research, Project No. 389-042, TR 8, Department of Geology, Columbia University, New York (1954).

[10] Shreve, R.L., "Statistical Law of Stream Numbers," Journal of Geology, *74*, 1 (1966).

A NOTE ON THE FLOW-SHOP PROBLEM WITHOUT INTERRUPTIONS IN JOB PROCESSING

Wlodzimierz Szwarc

School of Business Administration
University of Wisconsin-Milwaukee
Milwaukee, Wisconsin

ABSTRACT

This paper deals with a flow-shop problem where the n jobs are being processed uninterrupted by m machines. A comprehensive theory based on "an earliest starting time of a job" concept produced the most efficient solution method for a variety of optimization criteria. The paper also rectifies several known results in this area.

1. INTRODUCTION

Consider a flow shop problem (later called UFP) where n jobs are being processed without interruptions by m machines in the same technological order. This model was formulated and solved in 1972 by S. S. Reddi and C. V. Ramamoorthy [4] and by D. A. Wismer [6]. J. N. D. Gupta [1] generalized the method of [4] in 1976 providing a new theoretical derivation. The theory of [6] and [1], however, is more complicated than that of the original flowshop problem. This paper offers another framework utilizing the "earliest time when a job starts being processed" concept which makes the theory of the UFP simple. It shows that the approach based on the difference of earliest starting times of two consecutive jobs (see formula (5)) leads to the most efficient method for a variety of objective functions. The paper also rectifies several results and conclusions of the existing literature.

2. THE MODEL

We will adopt the following notations:

M_s — s-th machine, $s = 1, 2, \dots, m$,
r — job r, $r = 1, 2, \dots, n$,
t_{rs} — processing time of job r on M_s, $t_{rs} > 0$,
$T(P)$ — completion time of permutation $P = p_1, p_2, \dots, p_n$ of numbers $1, 2, \dots, n$,
t_i^s — earliest time when M_s starts processing job p_i of sequence P.

For simplicity assume $t_i^1 = t_i$ and $t_1 = 0$. Then for each $k = 1, 2, \dots, m$

$$(1) \qquad t_j^k = t_j + \sum_{s=1}^{k-1} t_{p_j s},$$

and

(2) $$T(P) = t_n + \sum_{s=1}^{m} t_{p_n s}.$$

Since $t_{rs} > 0$, one can prove the following theorem of [1].

THEOREM 1: For a schedule to be feasible it is necessary that all machines process the jobs in the same order.

PROOF: Suppose machines M_k and M_{k+1} do not process the jobs in the same order. One can assume that job a follows immediately job b on M_{k+1}, while a precedes b on M_k. Since b is being operated without interruptions the processing of a will stop for at least $t_{bk} + t_{bk+1} > 0$ which violates the feasibility of the schedule, Qed.

Theorem 1 of [1] where $t_{rs} \geqslant 0$ is *not true*, however, as the authors of [3] have shown (by a counterexample).* According to (1) and (2) the t_j^k, and $T(P)$ can be calculated given the appropriate t_j. Define

(3) $$d(p_i,\ p_{i+1}) = t_{i+1} - t_i.$$

Then

(4) $$t_j = t_1 + \sum_{i=1}^{j-1} d(p_i,\ p_{i+1}) = \sum_{i=1}^{j-1} d(p_i,\ p_{i+1}).$$

To find $d(p_i,\ p_{i+1})$ we need the following theorem. Consider an arbitrary sequence P where $a = p_i$, $b = p_{i+1}$. Assume that the processing of job a starts at the earliest time t_i.

THEOREM 2:

(5) $$d(a,b) = \max_{1 \leqslant u \leqslant m} \left[\sum_{s=1}^{u} t_{as} - \sum_{s=1}^{u-1} t_{bs} \right].$$

PROOF: The problem of jobs a and b being processed uninterrupted on machines M_1, M_2, ... , M_m can be viewed as a classical two-machine flow-shop model where "machines" a and b handle sequence M_1, M_2, ... , M_m of "items" M_s in the same order a,b and where processing on the first "machine" a goes on uninterrupted. According to [2] the minimum waiting time of "machine" b that operates sequence M_1, M_2, ... , M_m is equal to the right hand side of (5). The processing on b will go on uninterrupted once b starts operating at $t_i + d(a,b)$ which is its earlier starting time t_{i+1} once $d(a,b)$ is minimum.

Observe that (5) holds for *every* sequence where job b immediately follows job a.

*The proof of this theorem (see [1], p. 236) is deficient due to the following gaps: a) it overlooked the fact that the waiting time for job *a may be* in *zero* whenever $t_{bm_1} = 0$ even though $t_{bm_2} > 0$ (and not $\geqslant t_{bm_1} + t_{bm_2}$) as stated [1], and b) different processing orders of m_1 and m_2 (say 12345 and 51234) do not necessarily assure an existence of a schedule where a immediately follows b on m_2 and b immediately follows a on m_1.

According to (2) and (4),

$$T(P) = t_n + \sum_{s=1}^{m} t_{p_n s} = \sum_{i=1}^{n-1} d(p_i, p_{i+1}) + \sum_{s=1}^{m} t_{p_n s}.$$

Hence, the UFP is equivalent to an $n+1$-city traveling salesman problem (TSP) where the distance cost matrix $C = \{c_{ab}\}$, $a,b = 0,1,\ldots,n$ is defined by

$$(6)\qquad \left.\begin{array}{ll} c_{0b} = 0, & \forall\ b \neq 0, \\ c_{ab} = d(a,b) & \forall\ a,b \neq 0,\ a \neq b, \\ c_{a0} = \sum_{s=1}^{m} t_{as}, & \forall\ a \neq 0, \\ c_{aa} = \infty, & \forall\ a. \end{array}\right\}$$

$P = p_1, p_2, \ldots, p_n$ is an optimal solution of the UFP if and only if $0, p_1, p_2, \ldots, p_n, 0$ is an optimal tour of the TSP.

3. IDLE TIME APPROACHES

Let $I_k(a,b)$ be the idle time between two consecutive jobs $a = p_i$ and $b = p_{i+1}$ on M_k. Then $I_k = t_b^k - t_a^k - t_{ak}$ or (see (1))

$$(7)\qquad I_k(a,b) = I_1(a,b) + \sum_{s=1}^{k-1} t_{bs} - \sum_{s=2}^{k} t_{as}.$$

Due to (5)

$$(8)\qquad I_1(a,b) = \max_{2 \leq u \leq m} \left\{ \sum_{s=2}^{u} t_{as} - \sum_{s=1}^{u-1} t_{bs},\ 0 \right\}.$$

While Reddi and Ramamoorthy [4] based their approach on $I_1(a,b)$ using (8), Gupta [1] preferred $I_k(a,b)$, $k \geq 2$, as defined by (7) and (8). To minimize $T(P)$* they solved an equivalent TSP where formula (6) for distance cost matrices C' of [4] and C'' of [1] reads**

$$(6')\qquad c'_{0b} = \sum_{r=1}^{n} t_{r1},\ c'_{ab} = I_1(a,b),\ c'_{a0} = \sum_{s=2}^{m} t_{as},\ c'_{aa} = \infty,$$

$$(6'')\qquad c''_{0b} = \sum_{s=1}^{m-1} t_{bs},\ c''_{ab} = I_m(a,b),\ c''_{a0} = 0,\qquad c''_{aa} = \infty.$$

*Gupta minimized $f(P) = T(P) - \sum_{i=1}^{n} t_{p_i m} = t_1^m + \sum_{i=1}^{n-1} I_m(p_i, p_{i+1}) = T(P) -$ Constant while Reddi-Ramamoorthy minimized $T(P) = \sum_{i=1}^{n} t_{p_i 1} + \sum_{i=1}^{n-1} I_1(p_i, p_{i+1}) + \sum_{s=2}^{m} t_{p_n s}$.

**Notice that $C' = C + \{u'_a + v'_b\}$, $C'' = C + \{u''_a + v''_b\}$ where C is defined by (6) and

$u'_0 = \sum_{r=1}^{n} t_{r1}$, $u'_a = -t_{a1}$, $1 \leq a \leq n$, $v'_b = 0$, $0 \leq b \leq n$,

$u''_0 = 0$ $v''_0 = 0$, $u''_a = -\sum_{s=1}^{m} t_{as}$, $1 \leq a \leq n$, $v''_b = \sum_{s=1}^{m-1} t_{bs}$, $1 \leq b \leq n$.

4. OTHER OPTIMIZATION CRITERIA

Let $W_k(P)$ be the waiting time of machine M_k (counted from $t_1 = 0$) that processes sequence P. Then

$$W_k(P) = t_1^k + \sum_{i=1}^{n-1} I_k(p_i, p_{i+1}).$$

Gupta [1] generalized the model of [4] using the following objective function.

$$f(P) = \sum_{k=2}^{m} w_k W_k(P). \tag{9}$$

Notice that for $w_m = 1$, and $w_k = 0$ otherwise, $f(P)$ and $T(P)$ differ by a constant (see first footnote on page 667).

One should point out that the summation in (9) must include term $w_1 W_1(P)$ since $W_1(P)$ is *not* fixed, contrary to what is claimed (see [1], p. 237). To show that $W_1(P)$ depends on P consider a 2-job, 2 machine UFP where $t_{11} = 1$, $t_{12} = 5$, $t_{21} = 6$, $t_{22} = 3$.

Then $W_1(1,2) = 0$ while $W_1(21) = 2$.

Another objective function is

$$f(P) = \sum_{k=1}^{m} w_k(t_n^k - t_1^k) = \sum_{k=1}^{m} w_k \left[\sum_{i=1}^{n} t_{p_i k}\right] + \sum_{k=1}^{m} \sum_{i=1}^{n-1} I_k(p_i, p_{i+1}) \tag{10}$$

which a weighted sum of time *involvement of all machines*. The first term of (10) being constant can be dropped without missing optimality. Each of these problems can be easily transformed into an equivalent TSP (using formulas similar to (11) of [1]. J. M. Van Deman and K. R. Baker [5] solved the USP with a mean flow time criterion. Since the completion time of job p_j on M_m, $t_j^m + t_{p_j m}$, is equal to $t_j + \sum_{s=1}^{m} t_{p_j s}$ (see (1)), the objective function is

$$f(P) = \frac{1}{n} \sum_{j=1}^{n} \left[t_j + \sum_{s=1}^{m} t_{p_j s}\right] = \frac{1}{n} \sum_{j=1}^{n} t_j + \text{Constant}.$$

Hence, one can instead minimize (see also (4))

$$z(P) = \sum_{j=1}^{n} t_j = \sum_{j=1}^{n} \sum_{i=1}^{j-1} d(p_i, p_{i+1}). \tag{11}$$

5. CONCLUSION

As we have learned, any objective function of the UFP can be expressed in terms of either $d(a, b)$, $I_1(a, b)$, or $I_k(a, b)$, as defined by (5), (8), and (7). Those formulas clearly indicate that $d(a, b)$ requires minimal computational effort, while $I_1(a, b)$ is second best.

One should obviously use $d(a, b)$ when dealing with the mean flow-time criterion (see (1)).

In all other cases, the efficiency of an approach should be measured in terms of time necessary to calculate the travel cost matrix of the respective TSP (see [1], p. 241). If the completion time $T(P)$ is being minimized, then according to (6), (6'), and (6") the $d(a, b)$ approach has an advantage over the remaining approaches since (5) is to be used $n(n - 1)$ times. The efficiency of the $d(a, b)$ approach diminishes, however, once Wismer's algorithm [6] is applied to determine the $d(a, b)$. Gupta [1] has shown that the approach based on $I_m(a, b)$ is more efficient than Wismer's method.*

Suppose the objective function explicitly depends on $I_k(a, b)$ as in Section 4. Taking advantage of (1) we can express $I_k(a, b) = t_b^k - t_a^k - t_{ak}$ as follows:

$$I_k(a, b) = d(a, b) + \sum_{s=1}^{k-1} t_{bs} - \sum_{s=1}^{k} t_{as}. \qquad (7')$$

Due to (5), formula (7') is computationally more efficient than (7).

REFERENCES

[1] Gupta, J.N.D., "Optimal Flowshop Schedules with no Intermediate Storage Space," Naval Research Logistics Quarterly, *23*, 235-243 (1976).

[2] Johnson, S.M., "Optimal Two- and Three-State Production Schedules with Setup Times Included," Naval Research Logistics Quarterly, *1*, 61-68 (1954).

[3] Panwalkar, S.S., M.L. Smith and C.R. Woollam, "Counterexamples to Optimal Permutation Schedules for Certain Flow-Shop Problems," Naval Research Logistics Quarterly, *28*, 339-340 (1981).

[4] Reddi, S.S. and C.V. Ramamoorthy, "On the Flow-Shop Sequencing Problem with No Wait in Process," Operational Research Quarterly, *23*, 323-331 (1972).

[5] Van Deman, J.M. and K.R. Baker, "Minimizing Mean Flowtime in the Flow Shop with No Intermediate Queues," AIIE Transactions, *6*, 28-34 (1974).

[6] Wismer, D.A., "Solution of the Flowshop-Scheduling Problem with No Intermediate Queues," Operations Research, *20*, 689-697 (1972).

*One can considerably improve the efficiency and clarity of Wismer's algorithm by reformulating its pivotal steps 3 and 4 as follows:

a. Find E_y^k and T_x^{k+1} from $E_y^k = T_y^{k-1} + r_y^{k-1}$ and $T_x^{k+1} = T_x^k + r_x^k$.

b. Is $E_y^k \geqslant T_x^{k+1}$? If not, set $d_{xy}^k = T_x^{k+1} - E_y^k = T_y^k = T_x^{k+1}$; go to 5.

If it is, set $d_{xy}^k = 0$ and $T_y^k = E_y^k$; go to 5.

This improvement still falls behind formula (5) in terms of efficiency.

A NOTE ON INTEGER SOLUTIONS TO LINEAR FRACTIONAL INTERVAL PROGRAMMING PROBLEMS BY A BRANCH & BOUND TECHNIQUE

S. C. Agrawal

Department of Mathematics
Deva Nagri Post-Graduate College
Meerut, India

Mam Chand

Department of Mathematics
Kisan Degree College, Simbhaoli
Ghaziabad, India

ABSTRACT

This paper provides a method for solving linear fractional interval programming problems in integers with the help of a branch and bound technique.

1. INTRODUCTION

This paper is concerned with describing a systematic procedure for solving a linear fractional interval programming problem with the additional condition that the variables are integers.

The problem is as follows:

$$\text{(1)} \qquad \text{maximize } Z = \left\{\frac{c'x + c_0}{d'x + d_0} = \frac{C(x)}{D(x)}\right\},$$

$$\text{subject to} \quad b^- \leqslant Ax \leqslant b^+,$$

x has integral components.

This is known as a fractional interval integer programming problem (FIIP). A linear fractional interval programming problem, abbreviated (FIP), is defined as:

$$\text{(2)} \qquad \text{maximize} \left\{\frac{c'x + c_0}{d'x + d_0} = \frac{C(x)}{D(x)}\right\},$$

$$\text{subject to} \quad b^- \leqslant Ax \leqslant b^+,$$

where c', c_0, d', d_0, b^-, b^+ and A are given.

The problem was introduced in [4] and solved explicitly in the feasible bounded case with A of full row rank, see also [2], [6].

The linear fractional programming problem in all generality was reduced by Charnes and Cooper [3] to, at most, a pair of ordinary linear programming problems, by adjoining a specified constraint to the given set of constraints.

A primal algorithm is given by A. Charnes, D. Granot and F. Granot [5] for solving the (FIP) problem directly and this utilizes the special structure of the interval constraints.

The purpose of this paper is to obtain an integer solution to the (FIP) problem and produces, after a finite number of iterations, an optimal integer solution to (FIIP).

2. PRELIMINARY RESULTS

Consider the fractional programming problem (FIP):

$$\text{maximize} \left\{ \frac{c'x + c_0}{d'x + d_0} = \frac{C(x)}{D(x)} \right\},$$

$$\text{subject to} \qquad b^- \leqslant Ax \leqslant b^+.$$

For the following, let b^-, $b^+ \in R^m$; c', d', $x \in R^n$; c_0, $d_0 \in R$ and A be a real mxn matrix. $N(A)$ denotes the null space of A, $R(A)$ the range space of $A, \perp$ the usual orthogonality relation in R^n, and R_r^{mxn} denotes the set of all mxn matrices of rank r.

To exclude trivial cases, the following assumptions are made:

ASSUMPTION 1: (FIP) is feasible, i.e.,

$$S = \{x \in R^n; b^- \leqslant Ax \leqslant b^+\} \neq \emptyset.$$

ASSUMPTION 2: $D(x) := d'x + d_0 > 0$ over S, the feasible region.

ASSUMPTION 3: There exists no $\lambda \in R$ such that

$$D(x) = \lambda C(x) := \lambda (c'x + c_0) \text{ for all } x \in S.$$

ASSUMPTION 4: $C \perp N(A)$ and $d \perp N(A)$, since a feasible (FIP) is unbounded if either $c \notin N(A)$, or $d \notin N(A)$ [4].

LEMMA 1: Let $A \in R_r^{mxn}$ and $D \in R_r^{rxn}$ be such that

$$R(D') = R(A') \tag{3}$$

then $AD' \in R_r^{mxr}$, i.e., AD' is of full column rank [9].

LEMMA 2 [5]: Let D be as in lemma 1 and suppose that (FIP) is given with $A \in R_r^{mxn}$ and $c \perp N(A)$, $d \perp N(A)$. Then, the optimal solution of (FIP) is:

$$D'y^* + (N(A)) \tag{4}$$

where y^* is any optimal solution of

$$(5) \qquad \text{maximize } \frac{c'D'y + c_0}{d'D'y + d_0}$$

$$\text{subject to} \quad b^- \leqslant A\ D'y \leqslant b^+.$$

PROOF: Since $R(A') = N(A)^{\perp}$ and since $c \in N(A)^{\perp}$, $d \in N(A)^{\perp}$, it follows that the optimal solution to (FIP) is of the form

$$(6) \qquad x^* + N(A).$$

Where x^* is an optimal solution to

$$(7) \qquad \text{maximize } \frac{c'x + c_0}{d'x + d_0},$$

$$\text{subject to} \quad b^- \leqslant Ax \leqslant b^+,$$

$$x \in R(A').$$

But, since $R(A') = R(D')$ and $x \in R(A')$ can be equivalently written as:

$$x = D'y, \qquad y \in R^r.$$

Substituting $x = D'y$ in (7) results in the equivalent problem, which completes the proof.

Without loss of generality it can be assumed, that there exists an $x \in S$ such that $D(x) > 0$.

Following Martos [8], every feasible point x to (FIP) is classified according to the table below:

x	Type 1	Type 2
"good" point	$D(x) > 0$	$D(x) = 0,\ C(x) < 0$
"bad" point	$D(x) < 0$	$D(x) = 0,\ C(x) > 0$
"singular" point	$D(x) = 0 = C(x)$	—

A complete analysis of (FIP) in all generality and generation of an optimal solution to (FIP), if one exists, is given by Charnes and others [5].

3. ALGORITHM

Let us assume that (FIP), given in (2), is feasible, and let $c \perp N(A)$, $d \perp N(A)$ and A be of full column rank representation (see Lemma 2) and $D(x) > 0$ on S.

From Martos [8], we recall:

THEOREM 1: If $S \neq \emptyset$ and the above assumptions are satisfied, then the fractional function attains a finite maximum on S, which is taken on at least one extreme point.

Let x^* be a feasible extreme point of S, and B a basis for the rows of A which includes all the rows of the linearly independent constraints satisfied as equalities at x^*, and N be the completion of B to A. Let b_B^-, b_N^-, b_B^+, b_N^+ be the partitions of the vectors b^- and b^+ which correspond to the partition of A to B and N, respectively.

(8) Let $y^* = Bx^*$,

and

$$z_i(y) = (c'B^{-1})_i\,(d_0 + d'B^{-1}y) = (d'B^{-1})_i\,(c_0 + c'B^{-1}y) \tag{9}$$

LEMMA 3: x^* is an optimal solution to (FIP) if

$$\left.\begin{array}{ll} z_i(y^*) \leqslant 0 & \text{whenever } y_i^* = b_i^- \\ z_i(y^*) \geqslant 0 & \text{whenever } y_i^* = b_i^+ \end{array}\right\} [5]. \tag{10}$$

An algorithm given in [5] shows that, after a known number of iterations, an extreme point solution is obtained. If this extreme point satisfies the optimality criterion (10), then the algorithm terminates with an optimal solution to (FIP) given by $x^{opt} = B^{-1}y$. If, however, the optimality criterion is not satisfied, the algorithm proceeds along adjacent extreme points while improving the value of the objective function until an optimal extreme point to (FIP) is produced.

We solve (2) as in [5] and thus we obtain the maximum value of the objective function as Δ, say. If all the components of x are also all integers, then this is obviously the required optimum solution. If the components of x are not all integers, we obtain one after another more restrictive upper bounds Δ_1, Δ_2, ..., Δ_l in the same way as Land and Doig [7] has done for linear objective functions.

It is possible to transfer the problem (1) to an equivalent linear programming problem and then to apply existing methods for solving the equivalent problem in integers. However, since such a transformation will increase the effective size of the problem and will destroy the special structure of the constraints, we apply Land and Doig method to problem (1) to obtain an integer solution. It maintains the original two-sided constraints structure and takes advantage of this special structure during the solution procedure and thus will be more efficient for problems which occur in (IP) form rather than the equivalent (LP) codes.

Let the value of any component of x, say x_p, be α (nonintegral) for the maximum value Δ of the objective function. Now, x_p is forced to take an integral value and hence is decreased to at least α^0 or increased to at least $\alpha^0 + 1$, where α^0 denotes the largest integer less than or equal to α.

We now solve the fractional interval programming problems by substituting $x_p = \alpha^0$, and then we solve it by substituting $x_p = \alpha^0 + 1$ by the same method [5]. Let the two values of the objective function thus obtained be Δ' and Δ'', respectively. However, if one of the above two problems, say the fractional interval program with the condition $x_p = \alpha^0$ is not feasible, Δ' does not exist. It implies that the value of x_p which only satisfies $x_p \geqslant \alpha^0 + 1$, need be considered in future discussion. If neither Δ' nor Δ'' exists, x_p cannot be constrained to an integral value and the problem possesses no feasible solution.

We now find $\Delta_1 = \max(\Delta', \Delta'')$. For the objective function having Δ_1 as one of its values, let the value of the variable x_p be β (integral), and also suppose all the other components of x are still not integers. To find the second best solution, determine the maximum values of the fractional interval program with $x_p = \beta - 1$ and then, with $x_p = \beta + 1$. (One of these values has already been obtained as Δ' or Δ''.) Also, as Δ_1 is not the required solution, a new variable (say, x_q) is chosen from those that are not integral at this stage. Let x_q be equal to γ (nonintegral), say. Hence, two more interval programs are solved, viz., the fractional interval program with $x_p = \beta$, $x_q = \gamma^0$; and then, with $x_p = \beta$, $x_q = \gamma^0 + 1$. Let Δ_2 be the maximum value of the objective function amongst the values just obtained and those that are obtained by substituting $x_p = \beta - 1$ and $x_p = \beta + 1$. The whole argument can now be repeated with Δ_2 replacing Δ_1 as the current upper bound on the optimal value of the objective function.

Continuing the above process, a tree is formed each of whose vertices represents a known set of integer constraints (for example, the vertex associated with Δ_1 value represents $x_p = \beta$). A branch terminates if it reaches a vertex having a nonfeasible solution. Ultimately, either all branches are terminated in vertices having no feasible solutions, or else a vertex having the maximum value Δ_k, say, is reached for which all the components for x are integers. This must be the required optimum solution.

4. NUMERICAL EXAMPLE

For the sake of simplicity and easy understanding of the method, we shall now solve an example.

EXAMPLE:

$$\text{maximize} \quad Z = \frac{3x_1 + 20x_2 + 4x_3 + 4}{2x_1 + 15x_2 + 3x_3 + 3},$$

$$\text{subject to} \quad 8 \leqslant 3x_1 + x_2 + 3x_3 \leqslant 16,$$

$$2 \leqslant x_1 + 3x_2 \leqslant 9,$$

$$-1 \leqslant 3x_1 + 5x_2 + x_3 \leqslant 11,$$

$$3 \leqslant 2x_2 + 4x_3 \leqslant 10,$$

$$x_1,\ x_2,\ x_3 \text{ are integers.}$$

Thus, we have

$$A = \begin{bmatrix} 3 & 1 & 3 \\ 1 & 3 & 0 \\ 3 & 5 & 1 \\ 0 & 2 & 4 \end{bmatrix}, \quad b^- = \begin{bmatrix} 8 \\ 2 \\ -1 \\ 3 \end{bmatrix}, \quad b^+ = \begin{bmatrix} 16 \\ 9 \\ 11 \\ 10 \end{bmatrix}.$$

The problem when solved by primal algorithm method [5] gives

$$x = \begin{bmatrix} 4\frac{4}{19} \\ -\frac{14}{19} \\ 1\frac{7}{19} \end{bmatrix} \text{ and } Z = 1\frac{11}{17},$$

but it does not satisfy the integrality condition.

We determine the first integral values of min x_3, and max x_3. Here, the first integer value of min x_3 is 1. Hence, we maximize the given objective function, subject to the given constraints by substituting $x_3 = 1$. The problem becomes

$$\begin{aligned} &\text{maximize} && Z = \frac{3x_1 + 20x_2 + 8}{2x_1 + 15x_2 + 6}, \\ &\text{subject to} && 5 \leqslant 3x_1 + x_2 \leqslant 13, \\ &&& 2 \leqslant x_1 + 3x_2 \leqslant 9, \\ &&& -2 \leqslant 3x_1 + 5x_2 \leqslant 10, \\ &&& -1 \leqslant 2x_2 \leqslant 6. \end{aligned}$$

This problem when solved by [5] gives $x_1 = 3\frac{1}{2}$, $x_2 = -\frac{1}{2}$ and $Z = 1\frac{6}{11}$.

The first integer value of max x_3 is 2. This gives $x_1 = 3\frac{1}{2}$, $x_2 = -\frac{1}{2}$ and $Z = 1\frac{8}{17}$.

Further, min x_3 for its second integer value is $x_3 = 0$. With this additional substitution, we do not obtain any feasible solution.

The maximum value of Z upto this point is $1\frac{6}{11}$ with $x_1 = 3\frac{1}{2}$, $x_2 = -\frac{1}{2}$ and $x_3 = 1$. Thus, Δ_1 is given the value $1\frac{6}{11}$.

We now maximize the given objective function subject to the given constraints with $x_3 = 1$, $x_1 = 3$. The value of Z comes out to be $1\frac{5}{12}$ and $x_2 = 0$. This gives first integer solution. Further, the objective function subject to the given constraints with $x_3 = 1$, $x_1 = 4$ does not give any feasible solution.

Since $1\frac{8}{17} > 1\frac{5}{12}$, we solve the given problem with $x_3 = 2$, $x_1 = 3$, the second integer solution is obtained i.e., $x_1 = 3$, $x_2 = 0$, $x_3 = 2$ and $Z = 1\frac{2}{5}$.

Further, with $x_3 = 2$, $x_1 = 4$, the solution is infeasible.

Thus, the required optimal integer solution is $Z = 1\frac{5}{12}$, $x_1 = 3$, $x_2 = 0$, $x_3 = 1$.

ACKNOWLEDGMENT

We are indebted to the referee for valuable comments which helped us in the revision of this paper.

BIBLIOGRAPHY

[1] Ben-Israel, A., and A. Charnes, "An Explicit Solution of a Special Class of Linear Programming Problems," Operations Research, *16*, 1166-1175 (1968).

[2] Bühler, W., "A Note on Fractional Interval Programming," Operations Research, *19*, 29-36 (1975).

[3] Charnes, A., and W.W. Cooper, "Programming with Linear Fractional Functionals," Naval Research Logistics Quarterly, *9*, 181-186 (1962).

[4] Charnes, A., and W.W. Cooper, "An Explicit General Solution in Linear Fractional Programming," Naval Research Logistics Quarterly, *20*, 449-467 (1973).

[5] Charnes, A., D. Granot and F. Granot, "On Solving Linear Fractional Interval Programming Problems," Faculty of Commerce and Business Administration, Working paper No. 358, University of British Columbia, Vancouver (Oct. 1975).

[6] Charnes, A., D. Granot and F. Granot, "A Note on Explicit Solution in Linear Fractional Programming," Naval Research Logistics Quarterly, *23*, 161-167 (1976).

[7] Land, A. and A. Doig, "An Automatic Method of Solving Discrete Programming Problems," Econometrica, *28*, 497-520 (1960).

[8] Martos, B., "Hyperbolic Programming," translated by A. and V. Whinston, Naval Research Logistics Quarterly, *11*, 135-155 (1964).

[9] Martos, B., "The Direct Power of Adjacent Vertex Programming Methods," Management Science, *12*, 241-252 (1965).

[10] Robers, P.D., "Interval Linear Programming," Ph.D. Dissertation, Northwestern University, Evanston, IL (1968).

[11] Zionts, S., "Programming with Linear Fractional Functionals," Naval Research Logistics Quarterly, *15*, 449-452 (1968).

A NOTE ON SOJOURN TIMES IN M/G/1 QUEUES WITH INSTANTANEOUS, BERNOULLI FEEDBACK*

Ralph L. Disney

Department of Industrial Engineering and Operations Research
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

ABSTRACT

Queueing systems which include the possibility for a customer to return to the same server for additional service are called queueing systems with feedback. Such systems occur in computer networks for example. In these systems a chosen customer will wait in the queue, be serviced and then, with probability p, return to wait again, be serviced again and continue this process until, with probability $(1 - p) = q$, it departs the system never to return. The time of waiting plus service time, the nth time the customer goes through, we will call his nth sojourn time. The (random) sum of these sojourn times we will call the total sojourn time (abbreviated, sojourn time when there is no confusion which sojourn time we are talking about). In this paper we study the total sojourn time in a queueing system with feedback. We give the details for M/G/1 queues in which the decision to feedback or not is a Bernoulli process. While the details of the computations can be more difficult, the structure of the sojourn time process is unchanged for the M/G/1 queue with a more general decision process as will be shown. We assume the reader is familiar with Disney, McNickle and Simon [1].

1. INTRODUCTION

One of the major, largely unsolved problems in queueing systems with feedback, is the sojourn time problem. Upon entry to the system, a customer spends some time waiting and being served. We call this total the customer's sojourn time on his first pass through the server. At the conclusion of this time, the customer immediately returns for more service, with probability p, or departs never to return. In the latter case the customer's total sojourn time and it's sojourn time on the first pass are equal. In the former case the customer immediately experiences another sojourn time on it's second pass. Thus, in general, the total sojourn time is a (random) sum of the sojourn times on each pass through the server. However, the number of passes through the server is a random variable. Under the assumption that the decision to feed back or not is a Bernoulli random variable, the number of passes through the server is then a geometrically distributed random variable.

*This research was supported jointly by NSF Grant ENG77-22757 and by the Office of Naval Research Contract N00014-77-C-0743 (NR042-296).

One can consider many such sojourn time problems depending on where the feedback unit is allowed to reenter the queue. If, for example, it returns to the head of the queue, one finds trivially that the total sojourn time is equivalent to the sojourn time of a customer in a queue without feedback. However, if the feedback reappears anywhere else in the line (e.g., the feedback goes to the end of the line as is the most common assumption), determining the total sojourn time distribution seems to be a formidable task. The problem is caused by the fact that it is the sum of the sojourn times that that customer spends each time he goes through the server and, these times are not independent of each other. We will study this total sojourn time problem.

Takacs [4] and Montazer-Haghighi [3] are the only papers that we are aware of that study this problem in detail. Takacs' results are for M/G/1 queues with an instantaneous Bernoulli feedback process, as are most of ours though our methods are different.

Montazer-Haghighi studies the M/M/m queue with instantaneous, Bernoulli feedback with particular emphasis on $m = 2$. He finds a transform for the sojourn time distribution in the steady state, and the first two moments of this distribution as well as the steady state queue length distribution.

2. THE PROBLEM AND NOTATION

We are concerned with sojourn times in M/G/1 queues with instantaneous, Bernoulli feedback. We assume the reader is familiar with Disney, McNickle and Simon [1].

Choose a customer, C. We will follow it through the system. Suppose that customer feeds back to the end of the queue, instantaneously K times in all. K is a geometric distributed random variable on $\{0, 1, 2, \ldots\}$ with parameter p. Each time the customer enters the queue, the queue discipline is first in-first out (FIFO). Then define:

N_0 = the number of customers ahead of C upon its initial arrival.

N_n = the number of customers left behind the nth time C leaves the server ($n = 1, 2, \ldots, K$).

T_n = the time at which C leaves the server for the nth time ($n = 1, 2, \ldots, K$).

$$Y_n = Y(T_n) = \begin{cases} 1, & \text{if } C \text{ feeds back at } T_n \\ 0, & \text{otherwise.} \end{cases}$$

$\{Y_n\}$ is a sequence of independent identically distributed random variables with $\Pr\{Y_n = 1\} = p$, $\Pr\{Y_n = 0\} = 1 - p = q$. That is $\{Y_n\}$ is a Bernoulli process.

$X_n = T_n - T_{n-1} =$ the sojourn time of C on its nth trip through the system ($n = 2, 3, \ldots, K$). That is, X_n is the sum of the time C spends waiting for service plus the time spent in service the nth time C enters the queue.

$X_1 = T_1 =$ the sojourn time of C on it's first trip through the server.

$T^f = X_1 + X_2 + \ldots + X_K$

$B_n(y)$ = number of exogenous arrivals to the queue during the nth sojourn time (of length y) of $C (n = 1, 2, \ldots, K)$. $B_n(y)$ is a Poisson random variable (λy) for each n.

Given N_{n-1},

C_n = the number of customers ahead of C at the start of it's nth pass through the system who feedback. C_n is a binomial random variable, with parameters p, N_{n-1}.

3. RESULTS

A point often made concerning the queue length process for the M/G/1 queue with $\{Y_n\}$ a Bernoulli process is that it is equivalent to a queue without feedback in the sense that there always exists an M/G/1 queue without feedback whose limiting queue length distribution is equal to that for the M/G/1 queue with Bernoulli feedback. The question then arises as to whether there is an M/G/1 queue without feedback whose limiting sojourn time distribution is equal to that of the M/G/1 queue with Bernoulli feedback. Lemma 1 is a partial answer to the question.

LEMMA 1: If the feedback unit goes to the tail of the line and if the M/G/1 queue without feedback and the M/G/1 queue with Bernoulli feedback are to have the same first three moments for their respective total service time distributions, then there is no M/G/1 queue without feedback whose limiting total sojourn time distribution is equal to that of the M/G/1 queue with Bernoulli feedback.

PROOF: Let T^w be the sojourn time for a *M/G/1 queue without feedback*. Let α_n be the nth moment about 0 of the service time (for one pass through the server) of the feedback queue. Let τ_n be the nth moment about 0 for the total service time in the queue without feedback. If the total service times of the two queues are to have the same first three moments then it is a simple exercise to show

$$\tau_1 = \frac{\alpha_1}{q},$$

$$\tau_2 = \frac{\alpha_2}{q} + \frac{2(1-q)\alpha_1^2}{q^2},$$

$$\tau_3 = \frac{\alpha_3}{q} + \frac{6(1-q)\alpha_1\alpha_2}{q^2} + \frac{6(1-q)^2\alpha_1^3}{q^3}.$$

Then using the expected sojourn time and the second moment of the sojourn time for the queue with feedback given by Takacs [4] and the well known results (for example, see Kleinrock [2]; Section 5.7) of corresponding moments for the M/G/1 queue without feedback one obtains

$$E(T^w) = E(T^f),$$

$$E((T^w)^2) \neq E((T^f)^2)$$

unless $q = 1$ (i.e., there is no feedback). □

The following results structure the problem as a Markov renewal process for the M/G/1 queue with $\{Y_n\}$ a Bernoulli process.

THEOREM 1: $\{N_n, X_n\}$ is a delayed Markov renewal process for each $n \leqslant K$.

PROOF: For $K = k$. One needs only note that when C first enters the queue, it finds N_0 customers already in line and, for any FIFO, M/G/1 queue X_1 depends only on N_0. If C does not feedback, X_1 is its total sojourn time. However, if C feeds back it will encounter all of those customers to arrive during X_1 and, additionally, those among the N_0 who feedback. But, the number ahead of C if it feeds back, N_1, depends only on X_1 and N_0. Since these results are true for any $n > 1$, $\{.V_n,\ X_n\}$ has the Markov renewal property. In the M/G/1 case this process is a delayed Markov renewal process since X_1 will not in general have the distribution of X_n for $n = 2, 3, \ldots$. □

Let $H_c(x)$ be the remaining service time of the customer in service (if any) at the arrival of C. Let $H^i(x)$ be the i-fold convolution of H with itself. Then we have

COROLLARY 1: The transition functions for the $\{N_n,\ X_n\}$ process are given by

$$D_{ij}(x) = P(N_1 = j,\ X_1 \leqslant x \mid N_0 = i)$$
$$= \begin{cases} \int_0^x H(dy)A_{0j}(y), & \text{if } i = 0, \\ \int_0^x (H_c * H^i)(dy)A_{ij}(y), & \text{if } i = 1, 2, \ldots \end{cases}$$

and

$$Q_{ij}(x) = P(N_n = j,\ X_n \leqslant x \mid N_{n-1} = i)$$
$$= \begin{cases} \int_0^x H(dy)A_{0j}(y), & \text{if } i = 0, \\ \int_0^x H^{i+1}(dy)A_{ij}(y), & \text{if } i = 1, 2, \ldots. \end{cases}$$

Here

$$A_{ij}(y) = \sum_{m=0}^{\min(i,j)} \binom{i}{m} p^m q^{i-m} \frac{(\lambda y)^{j-m}}{m!} e^{-\lambda y}.$$

PROOF: The result is made apparent as follows. Given i and C, the number feeding back is simply a Bernoulli random variable with parameters i,p. For a given C, and a given sojourn time the nth time C goes through, the number of new arrivals is a Poisson random variable. Furthermore, given C and i, the new arrival process and the feedback process are independent. Therefore, $A_{ij}(y)$ is simply the distribution function for the total number of inputs to the queue during C's nth pass through. Given there are i ahead of him on the pass, $H^{i+1}(T)$ or $H_c(T)$ is then the distribution function for the sojourn time of C on his nth pass. Then since

$$\Pr[N_n = j,\ X_n \leqslant x \mid N_{n-1} = i] = \Pr[X_n \leqslant x \mid N_{n-1} = i] \Pr[N_n = j \mid X_n \leqslant x,\ N_{n-1} = i]$$

the result follows easily. □

Then since $\{N_n,\ X_n\}$ is a Markov renewal process we have

COROLLARY 2:

$$Q_{ij}^{(k)}(y) = P(N_k = j,\ T_k^f \leqslant y \mid N_0 = i) =$$
$$= \int_0^x \sum_{m=1}^{\infty} D_{im}(dx) Q_{mj}^{(k-1)}(y - x).$$

PROOF: For $K = k$ these are the usual k step transition functions for a delayed Markov renewal process. □

If $\lambda < q\mu$ there exists a vector π satisfying

$$\pi = \pi Q(\infty).$$

THEOREM 2: For a M/G/1 queue with instantaneous Bernoulli feedback, in the steady state, the sojourn time for C is given by

$$P(T^f \leqslant y) = \pi \left[\sum_{k=1}^{\infty} Q_{ij}^{(k)}(y) p^{k-1} q \right] U,$$

where U is the column vector whose elements are all 1.

PROOF: Conditioned on C feeding back k times Corollary 2 gives the joint distribution of N_k and T_k^f conditioned on N_0. The result then follows by removing the condition for the number of feedbacks and the N_0 and finding the resulting marginal distribution for T^f. □

Thus the marginal sojourn time of C is given by Theorem 2. But notice that, in general, sojourn times of successive customers are not independent and therefore the sojourn time *process* is not specified by Theorem 2 alone. Notice also that the Bernoulli process $\{Y_n\}$ plays a minor role. The results can be generalized at least to the case

$$P(Y_n = u \mid Y_{n-1} = \text{v},\ N_n - N_{n-1} = i,\ S_n = y).$$

Thus, in principle, the sojourn time problem for M/G/1 queues with feedback is solved for a rather large class of $\{Y_n\}$ processes.

There is an interesting question in these results. We know from Lemma 1 that there is no M/G/1 queue without feedback whose sojourn time distribution is equal to the M/G/1 queue with Bernoulli feedback if the two service times are to have the same first three moments. If we relax this moment requirement, the problem can be exposed as follows: Is there any M/G/1 queue without feedback whose sojourn time distribution for a given customer is equal to that sojourn time for a given M/G/1 queue with Bernoulli feedback?

REFERENCES

[1] Disney, R.L., D.C. McNickle, and B. Simon, "The M/G/1 Queue with Instantaneous Bernoulli Feedback," Naval Research Logistics Quarterly, *27*, 635-644 (1980).

[2] Kleinrock, L., *Queueing Systems, Vol. 1: Theory,* (John Wiley and Sons, New York, 1975).

[3] Montazer-Haghighi, A., "Many Server Queueing Systems with Feedback," 228-249, *Proceedings of the Eighth National Mathematics Conference*, Arya-Mehr University of Technology, Tehran, Iran (1977).

[4] Takacs, L., "A Single Server Queue with Feedback," Bell System Technical Journal, *42*, 505-519 (1963).

A NOTE ON TWO IFR SYSTEMS

Zvi Schechner

Department of Industrial Engineering and Operations Research
Columbia University
New York, N.Y.

ABSTRACT

We present probabilistic proofs for the following two facts:

(i) A k out of n system of i.i.d (independent identically distributed). IFR (increasing failure rate) components has an IFR life distribution.

(ii) A compound Poisson process with nonnegative i.i.d jumps with PF_2 distribution is IFR.

INTRODUCTION

A nonnegative random variable T is said to possess an IFR distribution if for any $a > 0$, $P(T > t + a | T > t)$ is nonincreasing in t. In words, given $\{T > t\}$, T is stochastically decreasing in t. This notion of aging is intuitively well understood and plays an important role in the theory of reliability. In this report we provide probabilistic proofs for the following known facts:

(i) A k out of n system of i.i.d. IFR components has an IFR life distribution

(ii) A compound Poisson process with nonnegative i.i.d. jumps with PF_2 distribution is IFR.

These theorems have been proven before, (i) in [2] and (ii) in [4]. In both cases the approach was purely analytical. By using a probabilistic approach we are able to provide better insight and intuition.

1. k-OUT-OF-n SYSTEM

Consider an$-n$ component coherent system ϕ. The components' lifelength $T_1, T_2, \ldots, T_n$ are assumed to be i.i.d. random variables having distribution $F(\bar{F} = 1 - F)$ and density $F' = f$ (for further discussion of coherent systems consult Barlow-Proschan [3]). Let $X_i(t)$, $i = 1, \ldots, n$ be 0 or 1 according to whether $\{T_i \leqslant t\}$ or $\{T_i > t\}$, respectively, and let $\underline{X}(t) = (X_1(t), \ldots, X_n(t))$ be the state vector at time t. Let $N_t = \sum_{i=1}^{n} X_i(t)$ denote the number of functioning components at t. Then for an arbitrary coherent structure we have:

LEMMA 1:

For any k, $(k = 0, 1, \ldots, n)$

$$E(N_t \mid N_t \geqslant k, \phi(\underline{X}(t)) = 1) \geqslant E(N_t \mid \phi(\underline{X}(t)) = 1).$$

PROOF:

Check that for $j = 1, \ldots, n$

$$P(N_t \geqslant j \mid N_t \geqslant k, \phi(\underline{X}(t)) = 1) \geqslant P(N_t \geqslant j \mid \phi(\underline{X}(t)) = 1).$$

LEMMA 2:

$$E(N_t \mid \phi(\underline{X}(t)) = 1) = \frac{\bar{F}(t) \sum_{j=1}^{n} h(1_j, \bar{F}(t))}{h(\bar{F}(t))}$$

where $h(\cdot)$ is the reliability function of ϕ and

$$h(1_j, \bar{F}(t)) = E(\phi(X_1(t), \ldots, X_{j-1}(t), 1, X_{j+1}(t), \ldots, X_n(t)).$$

PROOF:

$$E(N_t \mid \phi(\underline{X}(t) = 1) = \sum_{i=1}^{n} E[X_i(t) \mid \phi(\underline{X}(t)) = 1]$$

$$= \frac{\sum_{i=1}^{n} P(X_i(t) = 1, \phi(\underline{X}(t)) = 1)}{P(\phi(\underline{X}(t)) = 1)}$$

$$= \frac{\sum_{i=1}^{n} P(X_i(t) = 1, \phi(X_1(t), \ldots, X_{i-1}(t), 1, X_{i+1}(t), \ldots, X_n(t)) = 1)}{h(\bar{F}(t))}$$

and by the independent assumption this equals

$$\frac{\sum_{i=1}^{n} P(X_i(t) = 1)\, P(\phi(X_1(t), \ldots, X_{i-1}(t), 1, X_{i+1}(t), \ldots, X_n(t)) = 1)}{h(\bar{F}(t))}$$

$$= \frac{\bar{F}(t) \sum_{i=1}^{n} h(1_i, \bar{F}(t))}{h(\bar{F}(t))}.$$

LEMMA 3:

$$h'(p) = \sum_{i=1}^{n} h(1_i, p) - \sum_{i=1}^{n} h(0_i, p).$$

PROOF:

See Barlow and Proschan [3].

The following is the main theo.em of this section.

THEOREM 1:

Let ϕ be an arbitrary coherent structure, then for any k ($k = 0, \ldots, n$):

$$P(N_t \geqslant k \mid \phi(\underline{X}(t)) = 1) \text{ is nonincreasing in } t.$$

Thus, given the system is up at t, the number of functioning components of stochastically decreasing in t.

PROOF:

$P(N_t \geqslant k \mid \phi(\underline{X}(t)) = 1)$ can be written in the following way:

$$\frac{\sum_{j=k}^{n} c_j [\bar{F}(t)]^j [F(t)]^{n-j}}{h(\bar{F}(t))}$$

where c_j is the number of distinct ϕ-path sets of size j. Thus,

$$\frac{d}{dt} P(N_t \geqslant k \mid \phi(\underline{X}(t)) = 1)$$

$$= \frac{d}{dt} \frac{\sum_{j=k}^{n} c_j [\bar{F}(t)]^j [F(t)]^{n-j}}{h(\bar{F}(t))}$$

$$= \frac{1}{h^2(\bar{F}(t))} \left\{ \frac{h(\bar{F}(t)) f(t)}{\bar{F}(t) F(t)} P(N_t \geqslant k, \phi(\underline{X}(t)) = 1) \times \right.$$

$$(E[N_t] - E[N_t \mid N_t \geqslant k, \phi(\underline{X}(t)) = 1])$$

$$\left. + f(t) h'(\bar{F}(t)) P(N_t \geqslant k, \phi(\underline{X}(t)) = 1) \right\}.$$

This is nonpositive for $t \geqslant 0$ iff

$$(*) \frac{h(\bar{F}(t))}{\bar{F}(t) F(t)} (E[N_t] - E[N_t \mid N_t \geqslant k, \phi(\underline{X}(t)) = 1]) + h'(\bar{F}(t)) \leqslant 0.$$

By Lemma 1,

$$(*) \leqslant \frac{h(\bar{F}(t))}{\bar{F}(t) F(t)} (E[N_t] - E[N_t \mid \phi(\underline{X}(t)) = 1]) + h'(\bar{F}(t)).$$

By Lemmas 2, 3:

$$= \frac{h(\bar{F}(t))}{\bar{F}(t)F(t)} \left[n\bar{F}(t) - \frac{\bar{F}(t) \sum_{i=1}^{n} h(1_i, \bar{F}(t))}{h(\bar{F}(t))} \right]$$

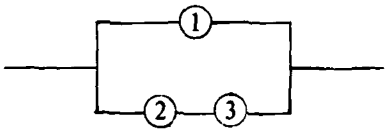
$$+ \sum_{i=1}^{n} h(1_i, \bar{F}(t)) - \sum_{i=1}^{n} h(0_i, \bar{F}(t))$$

$$= \frac{1}{F(t)} \left[nh(\bar{F}(t)) - \bar{F}(t) \sum_{i=1}^{n} h(1_i, \bar{F}(t)) - F(t) \sum_{i=1}^{n} h(0_i, \bar{F}(t)) \right]$$

$$= \frac{1}{F(t)} [nh(\bar{F}(t)) - nh(\bar{F}(t))] = 0.$$

Thus, $\frac{d}{dt} P(N_t \geq k \mid \phi(\underline{X}(t)) = 1) \leq 0$.

Note that even though the unconditional process N_t is decreasing stochastically (in fact, decreasing with probability 1, it is not always true for the process conditional on $\phi(\underline{X}(t)) = 1$. Indeed, for the nonidentical component cast this is not true. The following example illustrates it. Let the system be



Its structure function is $\phi(\underline{X}) = \max\{X_1, X_2 \cdot X_3\}$ and suppose the lifetimes are independent and distributed uniformlv on (0, 1) and (0, 2) (0, 2), respectively, then check

$$P(N_t \geq 2 \mid \phi(\underline{X}(t)) = 1) < 1 \quad \text{for } 0 < t < 1$$
$$= 1 \quad \text{for } 1 < t < 2.$$

DEFINITION:

A structure ϕ is called k-out-of-n iff $\phi(\underline{X}) = 1$ or 0 according to whether $\sum_{i=1}^{n} X_i \geq k$ or $< k$, respectively.

COROLLARY 1:

If the component lifetime is IFR, then the k-out-of-n system has an IFR distribution.

PROOF: Fix k, $(k = 1, 2, \ldots, n)$

We have to show that for $x > 0$

$$P(N_{t+x} \geq k \mid N_t \geq k) \text{ is decreasing in } t.$$

Let

$$S(r, t, x) = \sum_{j=k}^{r} \binom{r}{j} \left[\frac{\bar{F}(t+x)}{\bar{F}(t)}\right]^{j} \left[1 - \frac{\bar{F}(t+x)}{\bar{F}(t)}\right]^{r-j}$$

where $r \geqslant k$, $x > 0$ and $F(\cdot)$ is the component lifetime distribution function, which is assumed to be IFR. Thus for fixed $x > 0$, $S(r, t, x)$ is increasing in r and decreasing in t and

$$P(N_{t+x} \geqslant k \mid N_t \geqslant k) = E(S(N_t, t, x) \mid N_t \geqslant k).$$

By Theorem 1, given $N_t \geqslant k$, N_t is stochastically decreasing in t, which implies

$$E(S(N_t, t, x) \mid N_t \geqslant k) \text{ is decreasing in } t.$$

2. IFR SHOCK MODEL

Let $\{N_t : t \geqslant 0\}$ be a homogeneous Poisson process with intensity $\lambda > 0$ and let $X_1, X_2, \ldots$ be a sequence of i.i.d. nonnegative random variables independent of $\{N_t\}$ and distributed according to F. The process

$$\begin{aligned} Y(t) &= X_1 + \ldots + X_{N_t} && \text{if } N_t > 0 \\ &\quad 0 && \text{if } N_t = 0 \end{aligned}$$

is called a compound Poisson process. Esary, Marshall and Proschan [4] studied this process and derived, among other things, conditions which make the process IFR. (A nonnegative process Y is IFR *iff for any* $a > 0$, $T_a = \inf\{t : Y_t \geqslant a\}$ has an IFR distribution.) Their approach is purely analytical and they derive the conditions using total positivity theory. We start with the following theorem:

THEOREM 3:

For any $k \geqslant 0$, $c \geqslant 0$:

$$P(N_t \geqslant k \mid Y(t) \leqslant c) \text{ is increasing } t.$$

PROOF:

$$P(N_t \geqslant k \mid Y(t) \leqslant c) = \frac{\sum_{j=k}^{\infty} F^{*j}(c)\,(\lambda t)^j/j!}{\sum_{j=0}^{\infty} F^{*j}(c)\,(\lambda t)^j/j!}$$

where F^{*j} denotes the jth fold convolution of F. The proof is by induction on k.

For $k = 1$, check

$$\frac{\partial}{\partial t} \frac{\sum_{j=1}^{\infty} F^{*j}(c)\,(\lambda t)^j/j!}{1 + \sum_{j=1}^{\infty} F^{*j}(c)\,(\lambda t)^j/j!} \geqslant 0.$$

Now assume it holds for $k \leqslant n$, and to show that it holds for $k = n + 1$

$$\frac{\sum_{j=n+1}^{\infty} F^{*j}(c)\,(\lambda t)^j/j!}{\sum_{j=0}^{\infty} F^{*j}(c)\,(\lambda t)^j/j!} = \frac{\sum_{j=n}^{\infty}}{\sum_{j=0}^{\infty}} \times \frac{\sum_{j=n+1}^{\infty}}{\sum_{j=n}^{\infty}}.$$

By the induction hypothesis it suffices to show

$$\frac{\partial}{\partial t}\,\frac{\left(\sum_{j=n+1}^{\infty}\right)}{\left(\sum_{j=n}^{\infty}\right)} \geqslant 0$$

but

$$\frac{\partial}{\partial t}\,\frac{\left(\sum_{j=n+1}^{\infty}\right)}{\left(\sum_{j=n}^{\infty}\right)} = \frac{\left(\frac{\partial}{\partial t}\sum_{j=n+1}^{\infty}\right)\left(\sum_{j=n}^{\infty}\right) - \left(\frac{\partial}{\partial t}\sum_{j=n}^{\infty}\right)\left(\sum_{j=n+1}^{\infty}\right)}{\left(\sum_{j=n}^{\infty}\right)^2}.$$

It is easy to check that the numerator is nonnegative iff:

$$\frac{t}{n}\left(\frac{\partial}{\partial t}\sum_{j=n+1}^{\infty} F^{*j}(c)\,(\lambda t)^j/j!\right) - \left(\sum_{j=n+1}^{\infty} F^{*j}(c)\,(\lambda t)^j/j!\right) \geqslant 0$$

which is true since it is equal to

$$\begin{aligned} &[F^{*(n+1)}(\lambda t)^{n+1}/nn! \quad + F^{*(n+2)}(\lambda t)^{n+2}/n(n+1)! \quad + \ldots] \\ -\ &[F^{*(n+1)}(\lambda t)^{n+1}/(n+1)! \quad + F^{*(n+2)}(\lambda t)^{n+2}/(n+2)! \quad + \ldots] \geqslant 0. \end{aligned}$$

COROLLARY 2:

Theorem 3 holds even when $Y(t)$ is nonhomogeneous compound Poisson. We also need the following well known lemma.

LEMMA 4:

If the distribution function F is PF_2, i.e., if for any $x_1 < x_2$, $y_1 < y_2$:

$$\begin{vmatrix} F(x_1 - y_1) & F(x_1 - y_2) \\ F(x_2 - y_1) & F(x_2 - y_2) \end{vmatrix} \geqslant 0$$

then for any $x_1 < x_2$ and $n \geqslant 0$

$$F^{*n}(x_1)F^{*(n+1)}(x_2) \geqslant F^{*n}(x_2)F^{*(n+1)}(x_1).$$

PROOF:

See Esary-Marshall-Proschan [4], Theorem 4.9. The above states that if a sequence of i.i.d. random variables $X_1, X_2, \ldots$ have a PF_2 distribution F, then for any $x_1 < x_2$:

$P(X_1 + \ldots + X_n \leqslant x_1 \mid X_1 + \ldots + X_n \leqslant x_2)$ is decreasing in n.

COROLLARY 3:

If F is PF_2, then for $x_1 < x_2$

$$P(Y(t) \leqslant x_1 \mid Y(t) \leqslant x_2) \text{ is decreasing in } t.$$

PROOF:

$$\begin{aligned}
&P(Y(t) \leqslant x_1 \mid Y(t) \leqslant x_2)\\
&\quad = \sum_{n=0}^{\infty} P(Y(t) \leqslant x_1 \mid Y(t) \leqslant x_2,\ N_t = n) P(N_t = n \mid Y(t) \leqslant x_2)\\
&\quad = \sum_{n=0}^{\infty} P(X_1 + \ldots + X_n \leqslant x_1 \mid X_1 + \ldots + X_n \leqslant x_2) P(N_t = n \mid Y(t) \leqslant x_2)
\end{aligned}$$

given $Y(t) \leqslant x_2$, N_t is stochastically increasing in t and $P(X_1 + \ldots + X_n \leqslant x_1 \mid X_1 + \ldots + X_n \leqslant x_2)$ is decreasing in n and hence

$$\sum_{n=0}^{\infty} P(X_1 + \ldots + X_n \leqslant x_1 \mid X_1 + \ldots + X_n \leqslant x_2) P(N_t = n \mid Y(t) \leqslant x_2)$$

is decreasing in t.

COROLLARY 4:

If F is PF_2 then the compound Poisson process Y is IFR.

COROLLARY 5:

Suppose Y is nonhomogeneous compound Poisson with intensity $\lambda(\cdot)$, then if F is PF_2 and λ is increasing, Y is IFR process.

PROOF:

Have to show that for $x > 0$, $c \geqslant 0$,

$$P(Y(t+x) \leqslant c \mid Y(t) \leqslant c) \text{ is decreasing in } t.$$

Fix $x > 0$ and let $h(t,y) = \sum_{j=0}^{\infty} \dfrac{\left[\int_t^{t+x} \lambda(u)\,du\right]^j}{j!} F^{*j}(c-y)$ for $0 \leqslant y \leqslant c$. Now since λ is increasing in t, $h(t,y)$ is decreasing in t and y. But

$$P(Y(t+x) \leqslant c \mid Y(t) \leqslant c) = Eh((t, Y(t)) \mid Y(t) \leqslant c)$$

which is decreasing in t. A similar result concerning the nonhomogeneous compound Poisson was obtained by Abdel-Hameed-Proschan [1] using similar argument as in Esary-Marshall-Proschan [4].

REFERENCES

[1] Abdel-Hameed, M.S. and F. Proschan, "Nonstationary Shock Models," Stochastic Processes and Their Applications, *1*, 383-404 (1973).

[2] Barlow, R.E. and F. Proschan, *Mathematical Theory of Reliability* (John Wiley and Sons, New York, 1965).

[3] Barlow, R.E. and F. Proschan, Statistical Theory of Reliability and Life Testing (Holt, Rinehart and Winston, New York 1975).

[4] Esary, J.D., A.W. Marshall and F. Proschan, "Shock Models and Wear Processes," Annals of Probability, *1*, 627-649 (1973).

A NOTE ON THE TWO MACHINE JOB SHOP WITH EXPONENTIAL PROCESSING TIMES

Michael Pinedo

Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

Consider two machines, labeled 1 and 2. A set of tasks has to be processed first on machine 1 and after that on machine 2. A second set of tasks has to be processed first on machine 2 and after that on machine 1. All the processing times are exponentially distributed. We present a policy which minimizes the expected completion time of all tasks.

We consider two machines, labeled 1 and 2. There are n tasks which have to be processed first on machine 1 and after that on machine 2. This set of tasks will be called set A. Moreover, there are m tasks which have to be processed first on machine 2 and after that on machine 1. These tasks will be called set B. The processing time of task i, $i \in AUB$, on *machine 1 (2) is a random variable exponentially distributed with rate* $\lambda_i (\mu_i)$. Any of these rates may be infinite; that means that the corresponding processing time is zero. This model is usually called a Job Shop.

We are interested in a policy which minimizes the expected completion time of all tasks—the so-called makespan. A class of policies will be considered in which the decision-maker, at any time when a machine is freed, is allowed to review his policy and let his decision depend on the past history of the process.

The version of the above problem where the processing times of the tasks are deterministic has been treated by Jackson [2]. He presented a polynomial time algorithm to find the optimal schedule.

The special case where set B is empty is usually called a Flow Shop. So a Flow Shop is a Job Shop in which all tasks are required to pass through the successive machines in the same order. Bagga [1] treated the two-machine Flow Shop with exponential processing times, i.e., the Job Shop where set B is empty. We will give here a short description of his results as it will play a role in the proof of the main theorem in this note. As the order in which the tasks are processed on the second machine does not affect the makespan in the Flow Shop model, the sequence in which the tasks are processed on the two machines can be assumed to be the same. So the Flow Shop model is basically a sequencing problem. Let $j_1, \ldots, j_n$. a permutation of $1, \ldots, n$, denote the sequence in which the tasks go through the machines, i.e., at time $t = 0$ task j_1 starts being processed on machine 1; after it finishes its processing there, it starts on machine 2, while task j_2 starts on machine 1, etc. One would like to know the sequence which minimizes the expected makespan. Let $E(F(j_1, \ldots, j_n))$ denote the expected makespan when using sequence $j_1, \ldots, j_n$.

THEOREM 1 (Bagga):

(i) $$E(F(j_1, \ldots, j_{i-1}, j_i, j_{i+1}, j_{i+2}, \ldots, j_n)) \leqslant$$
$$E(F(j_1, \ldots, j_{i-1}, j_{i+1}, j_i, j_{i+2}, \ldots, j_n))$$

when

$$(\lambda_{j_i} - \mu_{j_i}) \geqslant (\lambda_{j_{i+1}} - \mu_{j_{i+1}})$$

(ii) Processing the tasks in decreasing order of $\lambda_i - \mu_i$ minimizes the expected makespan.

It is clear that (ii) is an immediate consequence of (i), but it is part (i) that we will need in the proof of Theorem 2. Theorem 1 tells us also that when the processing time of task i on machine 1 (2) is zero, i.e., $\lambda_i = \infty$ ($\mu_i = \infty$), it has to go first (last). If there are a number of tasks with zero processing times on machine 1, all these tasks have to precede all the others. The sequence in which these tasks go through machine 2 does not affect the makespan. A similar remark can be made if there is more than one task with zero processing time on *machine* 2.

The Job Shop model under consideration in this note can be viewed as a Markov Decision Process in continuous time. Define the decision moments to be the time epochs that a machine is freed. A policy instructs the decision maker to take a certain action at a decision moment (to start processing a certain task on the machine just freed) depending upon the state of the system at that moment. The state of the system at a decision moment is determined by the tasks which have been completed up to that moment and the task which is still being processed on the busy machine. In Theorem 2 we will determine the optimal policy, i.e., the policy which minimizes the expected makespan. We will use the following terminology: A "single" task denotes a task which only has to be processed on one of the machines; so we assume that a single task of set A (B) remains only to be processed on machine 2 (1). A "double" task denotes a task which still has to be processed on both machines in the prescribed order. Observe that a single task of set B can be considered as a double task of set A with zero processing time on machine 2.

THEOREM 2:

The optimal policy instructs the decision-maker, whenever machine 1 (2) is freed, to start processing of the remaining double tasks of set A (B) the one with the highest value of $\lambda - \mu$ ($\mu - \lambda$). If no double tasks of set A (B) remain, the decision-maker may start any one of the single tasks.

PROOF:

The proof consists of two parts. In the first part we compare two policies, which we will call π_1 and π_2. Suppose machine 1 is freed at time t. Both π_1 and π_2 will schedule the remaining double tasks of set B and the single tasks of set A, which finished their processing on machine 1 before time t, in the same sequence on machine 2; under π_1 and π_2 machine 2 has to process these tasks before it processes those tasks of set A which finish on machine 1 after time t. On machine 1 policy π_1 will schedule at time t first task 1 (with parameters λ_1 and μ_1) followed by task 2 (with parameters λ_2 and μ_2). Any of these parameters may be infinite. Policy π_2 will schedule first task 2 and then task 1. After finishing these two tasks, policies π_1 and π_2 will schedule the remaining tasks in the same sequence on machine 1. These remaining tasks include double tasks as well as single tasks, including the single tasks of set B which have finished their processing on machine 2 after time t. Call this sequence of tasks $j_1, \ldots, j_k$. Let x denote the time machine 2 still needs to finish the remaining double tasks of set B and the single tasks of set A which have finished their processing on machine 1 before time t. Let $X_1(X_2)$ denote the random processing time of task 1 (2) on machine 1. In case $X_1 + X_2 \leqslant x$ the makespans under π_1 and π_2 are clearly equal. So we only have to compare π_1 and π_2 in case $X_1 + X_2 > x$. We will make a distinction between two cases:

(i) Suppose $X_1 > x$ and we are using π_1. At time $t + x$ the problem reduces to a Flow Shop where the sequence used is $1, 2, j_1, \ldots, j_k$.

(ii) Suppose $X_1 < x$ and $X_1 + X_2 > x$. The problem reduces at time $t + x$ to the same Flow Shop with the difference that task 1 has finished its processing on machine 1 already.

Now let $E(J_1)$ $(E(J_2))$ denote the expected remaining time to finish all tasks on both machines under policy $\pi_1(\pi_2)$.

$$E(J_1 | X_1 + X_2 > x) \cdot P(X_1 + X_2 > x) =$$
$$e^{-\lambda_1 x}[x + E(F(1, 2, j_1, \ldots, j_k))] +$$
$$\int_{y=0}^{x} \lambda_1 e^{-\lambda_1 y} e^{-\lambda_2(x-y)}\, dy \cdot \left[x + E(F(1, 2, j_1, \ldots, j_k)) - \frac{1}{\lambda_1}\right]$$

and

$$E(J_2 | X_1 + X_2 > x) \cdot P(X_1 + X_2 > x) =$$
$$e^{-\lambda_2 x}[x + E(F(2, 1, j_1, \ldots, j_k))] +$$
$$\int_{y=0}^{x} \lambda_2 e^{-\lambda_2 y} e^{-\lambda_1(x-y)}\, dy \left[x + E(F(2, 1, j_1, \ldots, j_k)) - \frac{1}{\lambda_2}\right].$$

Clearly,

$$e^{-\lambda_1 x} + \int_{y=0}^{x} \lambda_1 e^{-\lambda_1 y} e^{-\lambda_2(x-y)}\, dy = e^{-\lambda_2 x} + \int_{y=0}^{x} \lambda_2 e^{-\lambda_2 y} e^{-\lambda_1(x-y)}\, dy$$

as both the l.h.s. and the r.h.s. are equal to $P(X_1 + X_2 > x)$. And also

$$\left[\int_{y=0}^{x} \lambda_1 e^{-\lambda_1 y} e^{-\lambda_2(x-y)}\, dy\right] \cdot \frac{1}{\lambda_1} = \left[\int_{y=0}^{x} \lambda_2 e^{-\lambda_2 y} e^{-\lambda_1(x-y)}\, dy\right] \cdot \frac{1}{\lambda_2}$$

as can be checked easily.

So

$$E(J_1|X_1 + X_2 > x) - E(J_2|X_1 + X_2 > x) =$$
$$E(F(1, 2, j_1, \ldots, j_k)) - E(F(2, 1, j_1, \ldots, j_k))$$

which according to Theorem 1 is positive when $\lambda_1 - \mu_1 < \lambda_2 - \mu_2$. This completes the first part of the proof.

In the second part of the proof we use the result of the first part to show the theorem. It is a well-known fact in the theory of Markov Decision Processes that a policy π^* is optimal, if when using π^* from any decision moment t and state onwards it results in a smaller expected makespan than acting at t *not* according to π^* but from the *next* decision moment onwards according to π^*.

Let π^* denote the policy described in the theorem and let π' denote a policy which acts differently at t and acts according to π^* from the next decision moment onwards. It is clear that the sequence resulting under π' can be transformed into the sequence under π^* through a number of adjacent pairwise switches, involving the tasks scheduled on the machine which was freed at time t. Each pairwise switch will cause a decrease in the expected makespan as was shown in the first part of the proof.

This completes the proof of the theorem.

REFERENCES

[1] Bagga, P.C., "n-Job, 2-Machine Sequencing Problem with Stochastic Service Times," Opsearch, *7*, 184-197 (1970).

[2] Jackson, J.R., "An Extension of Johnson's Results on Job Lot Scheduling," Naval Research Logistics Quarterly, *3*, 201-203 (1956).

INDEX TO VOLUME 28

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should bc sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.